Journal of Biomolecular NMR (2005) 31: 97–114

NMR structure determination of proteins supplemented by quantum chemical calculations: Detailed structure of the Ca^{2+} sites in the EGF34 fragment of protein S

Ya-Wen Hsiao[a], Torbjörn Drakenberg[b] & Ulf Ryde[a,]*
[a]*Department of Theoretical Chemistry and* [b]*Department of Biophysical Chemistry, Chemical Centre, Lund University, P.O. Box 124, S-221 00 Lund, Sweden*

Received 14 July 2004; Accepted 12 October 2004

Key words: Ca^{2+} sites, density functional calculations, EGF modules, NMR refinement, QM/MM methods

Abstract

We present and test two methods to use quantum chemical calculations to improve standard protein structure refinement by molecular dynamics simulations restrained to experimental NMR data. In the first, we replace the molecular mechanics force field (employed in standard refinement to supplement experimental data) for a site of interest by quantum chemical calculations. This way, we obtain an accurate description of the site, even if a molecular-mechanics force field does not exist for this site, or if there is little experimental information about the site. Moreover, the site may change its bonding during the refinement, which often is the case for metal sites. The second method is to extract a molecular mechanics potential for the site of interest from a quantum chemical geometry optimisation and frequency calculation. We apply both methods to the two Ca^{2+} sites in the epidermal growth factor-like domains 3 and 4 in the vitamin K-dependent protein S and compare them to various methods to treat these sites in standard refinement. We show that both methods perform well and have their advantages and disadvantages. We also show that the glutamate Ca^{2+} ligand is unlikely to bind in a bidentate mode, in contrast to the crystal structure of an EGF domain of factor IX.

Abbreviations: cbEGF – calcium-binding EGF domains; EGF – Epidermal growth factor; EGF34 – epidermal growth factor-like domains 3 and 4 in the vitamin K-dependent protein S; MM – molecular mechanics; NOE – nuclear Overhauser effect; QM – quantum mechanics; rMD – restrained molecular dynamics; SANI – susceptibility anisotropy.

Introduction

Nuclear magnetic resonance (NMR) and X-ray crystallography are the two major sources of structural information for large biomolecules, such as proteins. Both methods have in common that they do not directly give a three-dimensional image of the structure. Instead, the structure is determined by an involved process of interpreting the experimental raw data. In crystallography, the problem is that the phases of the reflections are unknown. Approximate phases can be obtained from related crystal structures or from heavy-metal derivatives, and they are then improved by repeated cycles of model building and refinement of the structure (Kleywegt and Jones, 1995). In NMR structure determination, the raw data consist mainly of a number of estimated distances between pairs of atoms, constraints in dihedral angles, and hydrogen bonds (Cavanagh et al., 1996). These are converted to a

*To whom correspondence should be addressed. E-mail: ulf.ryde@teokem.lu.se three-dimensional structure by the use of distance geometry methods or restrained molecular dynamics (rMD).

Both methods have also in common that the experimental data is usually supplemented by empirical chemical data, typically in the form of a molecular-mechanics force field, with terms for the ideal geometry of bonds, angles, dihedrals, planar groups, chirality, and non-bonded interactions. The force field is used to ensure that the bond lengths and angles are chemically reasonable and that aromatic systems are planar.

As a consequence, the quality of the resulting structures will depend on the force field used in the structure refinement (Kleywegt and Jones, 1998; Nilsson et al., 2003). For standard amino acids and nucleic acids, accurate target values for bond lengths and angles exist (Engh and Huber, 1991). However, for more unusual molecules, such as substrates, inhibitors, coenzymes, and metal centres, i.e., *hetero-compounds*, experimental data are often incomplete or less accurate (Kleywegt and Jones, 1998). In particular, force constants are normally not available and the force field has to be constructed by the experimentalist, a complicated and error-prone procedure.

A conceivable way to solve these problems is to replace the force field for the site of interest by more accurate quantum chemical calculations: Density functional calculations with a medium-sized basis set typically reproduce experimental bond lengths within 0.02 Å for organic molecules and 0.07 Å for bonds to metal ions (Jensen, 1999; Olsson and Ryde, 2001; Sigfridsson et al., 2001; Ryde and Nilsson, 2003a), making them more accurate than standard low- and medium-resolution crystal structures. We have recently developed such a method, *quantum refinement* (Ryde et al., 2002), in which we replace the empirical force field for a small part of the protein in a standard crystallographic refinement by quantum chemical calculations. We have shown that it works properly and that it can be used to locally improve crystal structures of hetero-compounds, e.g., inhibitors and metal sites (Ryde et al., 2002; Ryde and Nilsson, 2003a).

In this paper, we show that a similar method can also be used to locally improve the results of NMR structure determinations. For such structures, this method has the additional advantage that it can be employed for sites for which the experimental data give little information about the structure, e.g., for metal sites. Therefore, we test the method for two calcium sites in the epidermal growth factor-like domains 3 and 4 in the vitamin K-dependent protein S. This is an ideal test case, because the Ca^{2+} ion is known to have flexible geometric preferences, binding to 6–8 ligands with variable Ca–ligand bond lengths (da Silva and Williams, 1991), which makes it very hard to describe by standard molecular mechanics methods. We show that the method works properly and that we can obtain a much more detailed picture of the calcium sites than with standard methods. We also test another method to automatically obtain a molecular mechanics force field for a site of interest from a theoretical frequency calculation (Nilsson et al., 2003).

Methods

Hess2FF and ComQum-N

Standard NMR refinement is performed as a rMD annealing scheme with an energy function of the type

$$E_{\text{tot}} = E_{\text{MM}} + E_{\text{NMR}}, \qquad (1)$$

where E_{NMR} is the sum of all the NMR restraint energies (e.g., distance constraints based on nuclear Overhauser effect (NOE) data, dihedral constraints from the J couplings, hydrogen-bond restraints from amide proton exchange data, and susceptibility anisotropy (SANI) restraints from the residual dipolar couplings). E_{MM} is a standard MM energy function with bond, angle, dihedral angle, and non-bonded terms. Thus, the structure is obtained as a compromise between these two terms. The magnitude of the NMR restraints is arbitrary; therefore, each separate NMR term has a weight factor that determines the importance of this restraint relative to the MM restraints, which are in energy units.

QM data can be introduced into this energy in two ways. First, we can use QM calculations to construct MM parameters for the system of interest. There are many ways to perform such a parameterisation (Norrby and Liljefors, 1998). We have used a simple, fast, and automatic method, originally developed for the study of hetero-compounds in crystal structures

(Hess2FF) (Nilsson et al., 2003), but it is directly applicable also to NMR systems. It extracts the ideal bond lengths, angles, and dihedrals, as well as force constants from the Hessian matrix (i.e. the second derivative of the energy with respect to the coordinates) obtained from a QM optimised structure of a model of the interesting part of the protein. In this way, we get a more accurate description of the site of interest than a standard MM potential (if anyone exists at all), but there is still a risk that the site is not well determined at the MM level (which is likely for a Ca^{2+} site).

Alternatively and more accurately, we can replace the MM potential by a full QM calculation of the energy and the forces. Unfortunately, accurate QM methods can not yet be applied on a whole protein. Therefore, we have to restrict the QM calculations to a small but interesting part of the protein (e.g., a part that is poorly defined by the NMR restraints or not well described by the standard MM force field). This is done by partitioning the protein into two subsystems. System 1 consists of the site of interest that will be studied by QM methods, whereas system 2 contains the rest of the protein (and possibly parts of the surrounding solvent). We can then calculate the energy as:

$$E_{\mathrm{tot}} = E_{\mathrm{QM1}} - E_{\mathrm{MM1}} + E_{\mathrm{MM}} + E_{\mathrm{NMR}}, \qquad (2)$$

where E_{QM1} is the QM energy of the QM system, E_{MM1} is the MM energy of the same system, whereas E_{MM} and E_{NMR} have the same meaning as in Equation 1. The E_{MM1} term is needed to avoid double counting of energies in system 1 (i.e., to cancel the MM terms of system 1 in E_{MM}).

This energy expression is similar to that used in the standard combined QM and MM method (QM/MM), which is one of the most popular ways to treat proteins with QM methods (Ryde, 1996, 2003; Svensson et al., 1996; Monard and Merz, 1999; Mulholland, 2001):

$$E_{\mathrm{tot}} = E_{\mathrm{QM1}} - E_{\mathrm{MM1}} + E_{\mathrm{MM}}. \qquad (3)$$

Thus, our approach can equivalently be seen as a QM/MM method restrained to fit the NMR data.

Special attention is needed if there is a covalent bond between the QM system and the surrounding protein (a junction). This is a well-known problem in QM/MM methods and a simple and robust solution (Nicoll et al., 2001) is to truncate the QM system with hydrogen atoms, the positions of which are linearly related to the corresponding carbon atom in the protein (Ryde, 1996; Ryde et al., 2002). Of course, E_{MM1} is also calculated with these hydrogen atoms, so that artefacts introduced by the hydrogen truncation may cancel. The forces are the negative gradient of the energy in Equation 2, taking into account the relation between the H and C junction atoms using the chain rule (Maseras and Morokuma, 1995).

In the quantum chemical calculations, system 1 is represented by its wavefunction and the rest of the protein is modeled by point charges that polarise the QM system in a self-consistent manner. In the MM calculations, all atoms are described by the standard MM force field, but without any electrostatic interactions between the QM system and the surrounding protein, because they are already accounted for in the QM calculations. The electrostatic interactions within system 2 can be treated either in the QM calculations or in the MM calculations. In the former case, all interactions are considered, including interactions between bonded atoms (1–2, 1–3, and 1–4 interactions), which is normally not intended in the force field. In the latter case, the electrostatic interactions are treated in the intended way by the MM force field (typically ignoring 1–2 and 1–3 interactions and scaling down the 1–4 interactions by a constant factor). However, in order to obtain stable and reliable energies, a large cutoff distance need to be employed (ideally infinite), which is not always possible (many MM programs insist on calculating and storing a vector of all pairs of interactions to be included in the calculations, which may become too large to store in the internal memory). With an infinite cutoff, the two methods typically give very similar structures and relative energies.

We have implemented this energy expression (Equation 2) in a program, ComQum-N, by constructing an interface between the QM software Turbomole 5.6 (Ahlrichs et al., 2000) and the free and widely used software Crystallography & NMR System (CNS), version 1.1 (Brunger et al., 2000). The interface is based on our QM/MM software ComQum (Ryde, 1996; Ryde and

Olsson, 2001). The philosophy behind this approach is that there should be no change in the code of the QM and MM/NMR software. Instead, ComQum-N consists of a number of small programs which move information between the software, adding the forces and energies in a proper way.

An interface between Turbomole and CNS already exists in the quantum refinement software ComQum-X (Ryde et al., 2002). However, this had to be slightly modified to allow for the treatment of also hydrogen atoms (these are normally not resolved in X-ray crystal structures) and for the inclusion of point charges in QM calculations (an electrostatic model without any hydrogen atoms is meaningless, because no hydrogen bonds or solvation can be described).

Moreover, procedures and input files had to be developed for the CNS calculations with NMR-based restraints. These were based on the CNS standard input file anneal.inp (dynamical annealing with NMR restraints, using rMD). This file was modified in a few ways (as is detailed in the web page http://www.teokem.lu.se/~ulf/Methods/comqum_n.html): First, an extra coordinate file has to be read and written containing the fourth to eighth decimals (standard CNS reads coordinates in PDB format, i.e., with three decimals, which is not enough for proper convergence) (Ryde et al., 2002). Second, all dynamics sections were disabled, so that only the final minimisation is performed. The reason for this is that we want to perform only a local optimisation of the site of interest at the end of the NMR refinement. It would be a waste of computational power to perform QM calculations before the QM system was approximately assembled (i.e. before the QM groups are in proximity). In the energy and force calculations no optimisation is performed at all (the number of steps for the geometry optimisation is zeroed). Third, code was added to write out the energy terms and forces in separate files to be read by the ComQum-N interface. Fourth, the nonbonded force field was slightly modified, as will be discussed below.

The program flow of ComQum-N is shown in Scheme 1. It can be seen that ComQum-N consists of five small interface routines that move information (energy, forces, charges, and coordinates) forth and back between the QM and NMR programs. In each cycle of the geometry optimisation, the geometry of system 1 is relaxed by the total forces, keeping the geometry of system 2 fixed. Then, the geometry of system 2 is relaxed by an extensive (not only a single step as for system 1) energy minimisation with NMR restraints, keeping system 1 fixed, performed by CNS (but still without any rMD annealing). This way, we take advantage of the fact that the

```
Evaluate QM wavefunction of S1 including electrostatics of S2
Repeat
        Evaluate QM forces from S1 + electrostatics of S2 onto S1
        Evaluate CNS forces (from S1 and S2 onto S1), no electrostatics
        Add the QM and CNS forces
        Relax the geomentry of S1 using these forces*
        Change coordinates of S1 in CNS representation
        Calculate charges of S1
        Insert these charges into CNS representation
        Relax S2 by an NMR-restrained minimisation with S1 fixed
        Change coordinates of S2 in the QM representation (point charges)
        Evaluate QM wavefunction and energy of S1 including S2 electrostatics
        Evaluate CNS energy function
        Add energies
Until convergence
```

Scheme 1. The program flow in the ComQum-N program. Tasks performed by the QM program are shown in bold face, those performed by the CNS software are shown in italics, and those performed by the ComQum-N interface routines are underlined. S1 and S2 denotes systems 1 and 2.

*The geometry optimisation of S1 can be performed by any program, but for convenience, we have used the QM program.

NMR refinement is much faster than the QM calculation. This relaxation of system 2 is optional, but we see no reason not to perform it, because the NMR restraints will ensure that the structure is close to the true structure, and it will relieve strain that otherwise may build up between the QM system and the surrounding protein.

In practice, the five interface routines of ComQum-N are divided into four separate programs: one core routine that is independent of the QM and MM programs, two input routines that construct input files to this core routine from the particular QM and MM programs in text files with a standard format, and one output routine that reads the output from the core routine and write the data back into the specific QM or MM program (Ryde et al., 2002). In this way, the core ComQum procedure becomes independent of the actual programs used for the QM or NMR calculations), which makes porting to other programs easier and more lucid.

Applications on EGF34

In order to test the performance of Hess2FF and ComQum-N for NMR refinement, we have applied them to the two Ca^{2+}-binding sites in the epidermal growth factor-like domains 3 and 4 in the vitamin K-dependent protein S (EGF34). Modules homologous with epidermal growth factor (EGF) are common in extracellular proteins. They are found in a wide variety of animal proteins: connective tissue fibres, complement, blood coagulation and fibrinolytic proteins, as well as proteins involved in cell morphogenesis (Appella et al., 1988; Campbell and Bork, 1993; Stenflo et al., 2000). In fact, this module is the fourth largest protein family, present in 1% of the human proteins (Henikoff et al., 1997). The EGF modules are independently folding domains that usually consist of 40–50 amino acids and three disulphide bridges. They are often involved in protein–protein interactions that are Ca^{2+} dependent. Thus, a subset of the EGF motifs bind a Ca^{2+} ion in a conserved sequence. The Ca site is typically 6–7 coordinate, involving five residues from the protein (two back-bone amide groups and three Asp, Asn, Glu, or Gln residues) and a water molecule.

Some proteins contain many EGF modules in tandem repeats (Stenflo et al., 2000). Interestingly, the Ca affinity of such multimers is often higher than for the isolated modules. For example, in the pair of EGF modules 3 and 4 in protein S, module 3 has approximately the same Ca affinity as the isolated module 3, whereas the affinity of the fourth module is 8600 times larger in the pair than in the isolated module 4 (Stenberg et al., 1997a, 1997b).

Structures of many EGF-module protein are known, both from NMR and crystallographic studies (Stenflo et al., 2000). In particular, a 1.5 Å crystal structure of the EGF-like domains in human clotting factor IX has been presented (Rao et al., 1995). It shows two EGF domains, each binding a Ca^{2+} ion in a pentagonal bipyramidal manner (one carboxylate group binds bidentately).

The anticoagulant cofactor protein S has four EGF-like domains in tandem. Domains 2–4 are calcium-binding EGF domains (cbEGF). The Ca^{2+} binding to these domains are much stronger than to any other cbEGF studied so far. The smallest fragment with high Ca affinity is EGF34. We have used NMR to determine the three-dimensional structure of this protein fragment, hoping that it would reveal the reason to the high Ca^{2+} affinity. Standard multidimensional NMR has been used to estimate H–H distances and rMD has been used to obtain structures in agreement with the NMR distance restraints. Experimental details, as well as a discussion of the general structure, and its relation to the high Ca^{2+} affinity will be published elsewhere (Drakenberg et al., in preparation). This paper is restricted to a discussion of various methods to treat the Ca^{2+} sites.

The putative Cys–Cys bridges and Ca^{2+}-binding sites in EGF34 were identified from the consensus sequence of the EGF domains (Stenflo et al., 2000) and by comparison with the EGF domains in clotting factor IX (Rao et al., 1995): It was assumed that disulphide bridges are formed by the Cys residues 164–176, 171–185, 187–200, 206–215, 211–224, and 226–241. Likewise, we assumed that the two Ca^{2+}-binding sites are formed by residues Asp-160, Val-161, Glu-163, Asn-178, and Ile-179, as well as by Asp-202, Ile-203, Glu-205, Asn-217, and Tyr-218 (the second and fifth residues of each site bind by the back-bone amide oxygen atom, whereas the others bind by the side chains). In addition, both Ca^{2+} sites were assumed to bind one water molecule.

In the ComQum-N calculations, the protein is divided into two parts. System 1 consisted of $Ca(CH_3COO)_2\,(CH_3CONHCH_3)_2\,(CH_3CONH_2)$ (H_2O) (the same for both Ca^{2+} sites) and it was treated by quantum chemistry, whereas the rest of the protein (system 2) was treated entirely with standard NMR refinement methods. Thus, there were 12 junctions between the two systems: one each for the Asp, Glu, and Asn residues, and four for the two back bone groups (N and C^{β} for the residue containing the CO group and C and C^{β} for the next residue, containing the CO group). The second back-bone model in both sites contains an additional junction, because the next residue is Pro, giving a junction also for the C^{δ} atom.

The parameters for the junctions were exactly the same as for the original amino acids, except for the bonds to the junction hydrogen atom. This bond length was taken from the structure of the same fragment (CH_3COO^-, $CH_3CONHCH_3$, or CH_3CONH_2) optimised with the quantum chemical method. The force constant was calculated as the force constant of the original bond times the square of the quotient of the ideal original bond length (from the force field libraries) and the ideal bond length of the junction. This way, the forces of the bond with and without the junction will be equal. In addition, a few improper dihedral angles involving both QM and junction atoms had to be removed (they will not cancel between E_{MM1} and E_{MM}, owing to the movement of the junction atoms).

Computational details

The QM calculations were performed by density functional theory, using the Becke–Perdew-1986 exchange-correlation functional (BP86) (Perdew, 1986; Becke, 1988) and the standard medium-sized 6-31G* basis set for all atoms (Hehre et al., 1986). Only the five pure *d*-type functions were used. The calculations were sped up by expansion of the Coulomb interactions in auxiliary basis sets, the resolution-of-identity approximation (Eichkorn et al., 1995, 1997). These calculations were performed by Turbomole 5.6 (Ahlrichs et al., 2000). Such a method is known to give accurate and nearly converged geometries for metal-containing systems (Siegbahn and Blomberg, 2000; Ryde et al., 2001; Ryde and Nilsson, 2003a). The ComQum-N optimisations were performed in two steps: First, system 2 was allowed to relax and the full geometry was optimised until the change in energy between two iterations was below 10^{-4} Hartree and the maximum norm of the gradients was below 10^{-2} atomic units. Then, system 2 was fixed and the structure was further optimised with stricter convergence criteria, 10^{-6} Hartree (2.6 J/mol) and 10^{-3} atomic units (1.4 kJ/mol/Å, i.e., the default criteria in Turbomole). In the latter calculations, the maximum allowed movement of any atom (dqmax) was reduced to 0.03 atomic units (0.016 Å). Otherwise, extensive oscillations were often seen, although the same minimum and energy was normally obtained, but after many more iterations.

In some calculations, solvation effects were estimated using the continuum conductor-like screening model (COSMO) (Klamt and Schüürmann, 1993; Schäfer et al., 2000). These calculations were performed with default values for all parameters (implying a water-like probe molecule) and a dielectric constant (ε) of 80. For the generation of the cavity, a set of atomic radii have to be defined. We used the optimised COSMO radii in Turbomole (130, 200, 183, and 172 pm for H, C, N, and O, respectively, and 200 pm for Ca^{2+}) (Klamt et al., 1998).

The CNS calculations (Brunger et al., 2000) were performed with the standard protein-allhdg, water, and ion topology and parameter files. All protein atoms (including hydrogen atoms) were included in these calculations, but no water molecules (except the two Ca^{2+} ligands in some calculations). As mentioned above, the NMR calculations were performed with the file anneal.inp. In this file, we used the default values for most parameters. In particular the final weight factors for the NOE and dihedral constraints were 75 and 400 and the final force constant for the SANI restraints was 1.0. When the protein was relaxed, 10 cycles of final minimization consisting of 200 steps were run. In each standard rMD calculation with CNS, 200 structures were obtained from a preliminary structure, using random starting velocities.

In the calculations we included four types of NMR restraints, viz. 813 NOE distance restraints, 30 hydrogen bond restraints, 89 dihedral restraints, and 43 susceptibility anisotropy restraints. NOE restraints from methyl groups, degenerate methylene groups, and ambiguous assignments were averaged using the default sum mode. A final SANI force constant of 1.0 kcal/mol was used, because it resulted in calculated residual dipolar couplings matching the experimental ones within experimental errors. The SANI coefficients were optimised through a grid search: $a_0 = -0.0601$, $a_1 = -15$, and $a_2 = 0.35$.

Calibration of the QM method

We began the investigation with some calibrations of the QM method. To this end, we started from the Ca^{2+} site 1 in the crystal structure of the EGF-like domain in human clotting factor IX (Rao et al., 1995). The site was truncated in the same way as in the ComQum-N calculations (i.e., to $Ca(CH_3COO)_2(CH_3CONHCH_3)_2(CH_3CONH_2))$ and a water molecule was added (there is an obviously empty coordination site in the crystal structure). Then, this structure was optimised with a number of different methods and basis sets. In addition, we also optimised a number of structures starting from NMR structures of the two Ca^{2+} sites in EGF34. The most interesting results of these calculations are collected in Table S1.

It can be seen that the structure of the Ca^{2+} site is quite insensitive to the QM method. The Ca–O distances change by less than 0.01 Å when the method is changed from BP86 to B3LYP (Hertwig and Koch, 1997). An increase of the basis set from 6-31G* to the appreciably larger 6-311 + G(2d,2p) (Hehre et al., 1986) has a somewhat larger effect on the Ca–O distance, viz. a contraction by 0.02–0.05 Å. Inclusion of a continuum solvent (COSMO model) with a dielectric constant of 80 (similar to water) changes the Ca–O distances by 0–0.03 Å in a somewhat erratic manner. Different starting structures had a similar effect on the site: The individual Ca–O distances differ by up to 0.07 Å, but the average is essentially the same, 2.41 Å.

However, if the results are compared to the crystal structure of the EGF-like domain in human clotting factor IX, it can be seen that most of the optimised structures end up in six-coordinate Ca^{2+} sites with both carboxylate groups binding in a monodentate manner, whereas one of the carboxylate groups bind bidentately in the crystal structure. Some NMR structures also ended up in a bidentate structure and from these, it can be concluded that such a binding leads to an increase in the Ca–O distance of the carboxylate group from 2.34–2.41 to 2.51–2.55 Å. The average Ca–O distance also increases to 2.48–2.49 Å.

The change between mono- and bidentate binding of carboxylate groups (so-called carboxylate shifts) has been studied in several other systems, e.g., for Zn^{2+} and binuclear iron sites (Ryde, 1999; Torrent et al., 2001). From these studies, it is clear that there is only a minor energetic difference between mono- and bidentate binding. Therefore, the binding mode is mainly determined by what interactions the non-bonding carboxylate oxygen atom can form in the monodentate state. This is also observed in the present structures. In the monodentate sites, the two carboxylate groups form strong hydrogen bonds to the water ligand and to the amide hydrogen atoms of the Asn ligand. In the crystal structure, the latter hydrogen bond is retained, whereas the water molecule and the bidentate carboxylate group is exposed to solvent, where more ideal hydrogen bonds can be provided by the surrounding water molecules. Such effects can be simulated in the calculations by adding a water molecule in the second coordination sphere of the Ca^{2+} ion. This also led to bidentate structures (Table S1).

Considering that the crystal structure is bidentate, it is somewhat alarming that it has an average Ca–O distance of 2.40 Å, which is more similar to the monodentate than to the bidentate optimised structures. This is partly an effect of the missing water ligand, which has a 0.01–0.06 Å longer Ca–O distance than the average value in all monodentate structures. It is also partly an effect of the basis set (~0.03 Å). In order to check if the remaining difference is caused by the uncertainty in the crystal structure (typically at least ~0.1 Å (Fields et al., 1994; Cruickshank, 1999; Nilsson et al., 2003) or by systematic errors in the QM method, we also optimised the structure of Ca^{2+} in water. Experimentally, it is known that Ca^{2+} on average has eight ligands in water with a distance of 2.48 Å (Jalilehvand et al., 2001). A QM optimisation of $Ca(H_2O)_8^{2+}$ in D_{4d} symmetry (to avoid internal hydrogen bonds between the water molecules) gave a Ca–O distance of 2.47 Å with the BP86/6-31G* method and 2.49 Å with the 6-311 + (2d,2p) basis set, i.e. both in excellent agreement with experimental data. On the basis of these results, we decided to use the BP86/6-31G* method, which is much faster than with the larger basis set. However, it should then be kept in mind that this method overestimates the Ca–O distances by ~0.03 Å.

Results and discussion

There are several ways to treat a metal site in standard NMR structure determination by rMD simulations. First, the metal site can be totally ignored, using restraints only from the NMR raw data. This should give a structure as close as possible to the NMR data, but still bring the metal ligands in proximity, if the site is well-defined by the NMR data. However, it would not give any information about the binding mode of the ligands or the detailed structure of the metal site. Second, if the metal ligands are known beforehand, metal–ligand distances could be included as normal NOE distance restraints, using reasonable estimates of the bond lengths. This seems to be the most common method in standard structure determination (Downing et al., 1996; Saha et al., 2001; Wang et al., 2001; Tossavainen et al., 2003). This should give an improved structure of the metal site. Third, the metal–ligand interactions could be described by an MM potential like the surrounding protein, rather than by NOE restraints. This should give a much more accurate description of the details of the metal site. However, accurate MM potentials for metal ions are hard to construct, especially if the actual number and geometry of the ligands are not known or may change.

In the following sections, we will first test these three approaches for the two Ca^{2+} sites in EGF34 to see how the predicted structure of the Ca^{2+} site changes and how well the NMR restraints can be fulfilled. This is important to ensure that we do not enforce a site that violates the NMR raw data. It should be remembered that we do not have any direct experimental evidence of the actual Ca^{2+} ligands – the suggested ligands, mentioned above come simply from sequence alignments (Stenflo et al., 2000). We will then see if we can improve the structure of the Ca^{2+} site by the use of ComQum-N.

Refinement without any calcium restraints

We first performed an rMD annealing without any restraints for the Ca^{2+} ion (i.e., we employed only the standard NMR restraints and no quantum chemistry or MM potential for Ca^{2+}). Consequently, the structure contained neither Ca^{2+} nor any water molecules. In total, 200 different structures were obtained in this way using random starting velocities. The refinement consisted of a high-temperature dynamics, two slow-cool annealings, and a final minimisation. The dynamics and the first annealing simulations were performed in torsional space and a soft repulsion potential was used in all simulations.

The results of these calculations (Table S2) show that the two Ca^{2+} sites have poor geometries in all the structures obtained. For example, for site 1, only one structure has a maximum O–O distance (for the Ca ligands) shorter than 10 Å (8.8 Å; it is 4.8 Å in the crystal structure). The lowest maximum Ca–O distance for any structure is 6.4 Å, which is much larger than for a typical Ca^{2+} site (it is 2.6 Å in the crystal structure; we here assume that the Ca^{2+} ion resides at the midpoint between the two carbonyl ligand atoms – in the crystal structure of the EGF domain in factor IX, these two atoms are on opposite sides of the Ca^{2+} ion with an O–Ca–O angle of 172–175°). In addition, the total energy of the best Ca structures is very high. For site 2, the situation is somewhat better, with carbonyl distances almost half as long as for site 1. However, even the best structures have a maximum O–O distance of 5.6 Å, a maximum Ca–O distance of 3.1 Å, and high total energies.

The reason why site 2 is better defined by the NMR data than site 1 is that site 1 is at one end of the protein, whereas site 2 is in the middle of the protein, as can be seen in Figure 1. However, the main conclusion from this section is that the NMR data alone does not lead to any reasonable Ca^{2+} sites. In particular, it is impossible to speculate about any details of the sites, e.g., whether the carboxylic groups are bidentate or how many water molecules may coordinate to the site.

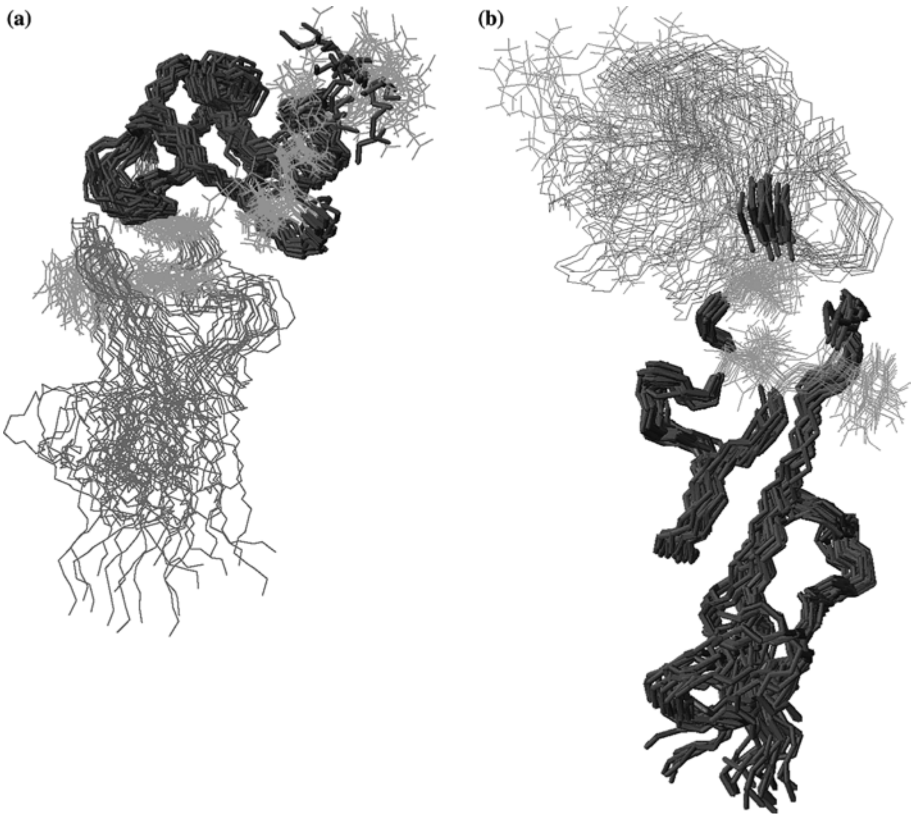

Figure 1. General structure of the EGF34 fragment with the ligands of the two Ca^{2+} sites emphasized in green. The best 20 structures are used, employing the data in Table S2 (no Ca^{2+} restraints). The left-hand side image was obtained by superimposing domain 3 (with Ca site 1), whereas the right-hand side image was obtained by superimposing domain 4 (with Ca site 2). The two domains are connected by a flexible hinge.

Finally, we can also note that there is an appreciable spread in the energies obtained for the various structures, both the total energies and the energies of the various NMR terms. For example, the average total and NMR energies of the best 20 structures are 654 and 92 kJ/mol, whereas the corresponding energies for the best structure are 529 and 72 kJ/mol. This is important to remember when judging the effect of various Ca^{2+} restraints.

Refinement with O–O restraints

Next, we tried to define the Ca^{2+} site by including a set of 10 O–O restraints between the five putative protein Ca^{2+} ligands for each site, defined in the same way as normal NOE restraints, with a flat-bottomed (between 3.0 and 5.1 Å) harmonic potential. Thus, we still did not include any Ca^{2+} ion or water molecules in the calculations.

Quite naturally, the O–O restraints ensure that all ligand atoms are relatively close in space, but in all structures, the carbonyl O–O distances are around the upper limit of the restraints, 5.1–5.2 Å and the maximum O–O distance is even longer 5.4–5.6 Å (Table S3). This is also reflected by the maximum Ca–O distance (the Ca^{2+} was inserted in the middle of the carbonyl O–O bond, as before), which is over 3.0 Å for all structures of site 1 and 2.8 Å for site 2. Thus, all Ca^{2+} sites still have effectively lost at least one ligand. Moreover, in most of the structures, some of the ligand oxygen atoms are not directed towards the putative centre of the Ca^{2+} site, as can be seen in Figure 2.

The energies of these structures are slightly larger than for those without any Ca^{2+}-related

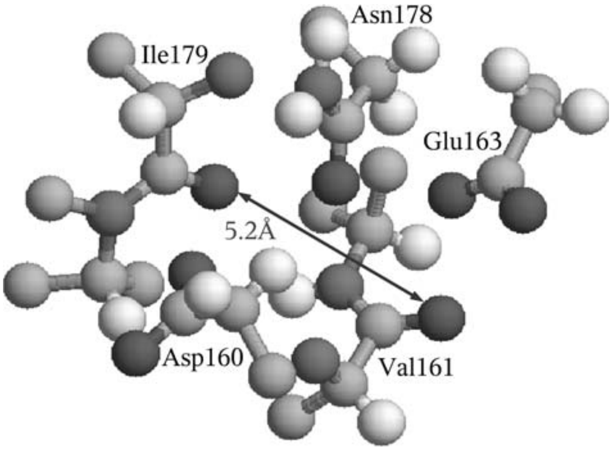

Figure 2. The structure of Ca^{2+} site 1, obtained with O–O restraints, showing that the orientation of some of the ligands are not proper for Ca binding.

restraints, e.g., by 205 kJ/mol for the average of the total energy for the 20 best structures. However, most of this difference comes from the MM energy; the difference in the total NMR energy is only 42 kJ/mol, originating mainly from the NOE term (36 kJ/mol). This can partly be explained by the additional O–O restraints for the Ca^{2+} sites, which are included in this term. It is a shortcoming of this method that the experimental data and empirical restraints are both mixed into the NOE term, so that it cannot be clearly determined how much the new restraints have affected the fit to the experimental data. However, it is notable that many of the best structures of the Ca^{2+} sites are also among the structures with the lowest energy. This shows that good sites are not unnatural.

The general structure of the protein obtained with these restraints (Figure S1) is similar to that obtained without any constraints. However, the Ca^{2+} sites, especially site 1, is somewhat better defined with these constraints. Thus, we can conclude that this seems to be a better method to obtain reasonable structures of the protein than without any Ca-restraints, but it still does not give any good geometries of the Ca^{2+} site.

Refinement with Ca–O NOE restraints

In order improve the structure of the Ca^{2+} site and get better starting points for the ComQum-N calculations, we decided to introduce the two Ca^{2+} ions and two water ligands in the refinement calculations. First, we tried to describe the Ca–O interaction with a flat-bottom potential, similar to the NOE restraints. This seems to be the most common way to treat a Ca^{2+} ion in NMR structures (Downing et al., 1996; Saha et al., 2001; Wang et al., 2001). The potential was zero between 2.0 and 3.0 Å and harmonic outside this range. However, this did not give any satisfactory results (Table S4): In all the obtained structures, the maximum Ca–O distance was 3.0 Å (the upper limit of the flat bottom). Thus, all sites have effectively lost at least one ligand and show little variation. Of course, we could have cured this problem by making the flat bottom tighter, but we would still only get what we put in (the upper limit), without any physical relevance of the results. Moreover, this approach, like the previous one, has the shortcoming of mixing up experimental data and the empirical potential, because the Ca^{2+} sites are described by NOE restraints and the corresponding energies will appear in the NMR term, rather than in the MM term.

Refinement with a bonded MM Ca–O potential

Therefore, we decided to use another approach, where the Ca–O interactions are described by a standard MM potential. The potential was obtained by the program Hess2FF (Nilsson et al., 2003) from QM vacuum optimisations and frequency calculations of the two Ca^{2+} sites in the EGF domain of factor IX. We used the structures in Table S1 and took force constants as the average of the similar interactions (i.e. for water, carboxylate, and carbonyl groups). The ideal bond lengths and force constants used are listed in Table S5. We decided to use only the bonded terms (i.e., no angle or dihedral restraints), because we did not want to bias the results towards any particular coordination number or geometry and also because the vacuum structure is somewhat distorted by interactions between the carboxylate groups and the methyl groups (an unavoidable vacuum effect).

The carboxylate groups pose a special problem because they can bind to Ca with either both or only one of the carboxylate oxygens. In the crystal structure, one of the carboxylate groups in each site binds in the bidentate mode, whereas the other binds monodentately. We decided to test both these possibilities in our calculations and therefore designed two sets of parameters, one for a monodentate site, based on the structure in the first row of Table S1, and the other bidentate, based on the sixth row in the same table (the calculation with an additional water molecule).

This approach gave excellent Ca^{2+} sites in all structures (Table S6). For both sites, the monodentate parameters gave an average maximum Ca–O bond length of 2.6–2.7 Å for all structures. The shortest maximum Ca–O bonds were 2.45 and 2.43 Å for the two sites, i.e., similar to what is found in the crystal structure. Likewise, the carbonyl O–O distances also show a quite restricted variation (averages 4.9–5.2 Å, slightly shorter for site 2 than for site 1; almost the same values were obtained for the maximum O–O distances).

The energies are similar to those obtained with the O–O restraints: The total energy of the best structure is 1 kJ/mol lower, but the average values of the total and the NMR energies in the 20 best structures are slightly higher. In particular, it seems that the dihedral terms have increased slightly.

The calculation with parameters for a bidentate binding of Glu-163/205 gave slightly worse results (Table S7): The average maximum Ca–O distances are ~0.2 Å longer in the bidentate structure, whereas in the best structures for each site, the difference is smaller (from the force field, the difference should be 0.09 Å). Likewise, the best and average energies are also larger for the bidentate site, by 40 kJ/mol for the total energy and by 11 kJ/mol for the NMR energy, this time originating mainly from the NOE term (average of the 20 best structures).

There is also the possibility that it is the Asp-160/202 residues that bind in a bidentate mode, although this is not observed in the crystal structure of the human clotting factor IX. Therefore, such a coordination was also tested. This gave actually slightly lower total and NMR energies than both the bidentate Glu sites and the monodentate sites (Table S8). For example, the average total and NMR energies of the 20 best structures were 880 and 121 kJ/mol for the bidentate Asp sites, whereas they were 946 and 144 kJ/mol for the monodentate sites. On the other hand, the Ca–O distances are appreciably longer in the bidentate Asp sites: The average maximum Ca–O distance among the 20 best structures is 2.78 and 2.90 Å for the two sites, whereas it is only 2.60 and 2.68 Å for the monodentate site. Once again, this is larger than what would be expected from the respectively equilibrium bond lengths in the force field (cf. Table S5).

In conclusion, bonded Ca–O terms of MM type, seems to be an excellent method to obtain reasonable structures for the Ca^{2+} sites. The resulting structure is at least as well determined as with the O–O restraints (Figure S2). The bidentate parameters for the Asp residues seem to give a slightly better site than the other two possibilities, but the NMR energies are not very different.

Calibration of the ComQum–N method

So far, we have only used the QM data to construct an MM potential of the Ca^{2+} site. Of course, much information is lost by this conversion and there is always the risk that errors are introduced, especially if the QM calculation is performed on a structure that is different from the final NMR structure. Moreover, a standard MM potential, such as the one in CNS does not allow the coordination number to change during the refinement. Therefore, we enforce a certain structure when we set up the MM potential, and this may not change during the refinement. Thus, we cannot model the dissociation of a ligand or the transition from mono- to bidentate binding of a carboxylate group.

All these problems can be avoided by using the QM calculations directly in the refinement, as in the ComQum-N method. However, we first have to test out the method and decide how it is optimally used. Therefore, we performed a number of test calculations, using various starting structures. We looked especially at four issues: the choice of repulsive parameters, the number of MM iterations, the weight of the NMR restraints, and the use of electrostatics in the QM and MM calculations.

In standard NMR refinements, CNS uses a soft repulsive potential, which allows atoms to go through each other. Unfortunately, this potential frequently led to failures in ComQum-N, because QM atoms ended up very close to MM atoms. For this reason, and also because it has been shown that such soft potentials, when employed also in the final minimisation of the refinement, may lead to poor structures in terms of the Ramachandran plots (Doreleijers et al., 1999), we decided to instead use the standard van der Waals (Lennard–Jones) potential of CNS (the default method for X-ray refinement). In addition, we used an infinite cut-off to avoid instabilities.

Second, we looked at the optimum number of iterations in the NMR-restrained minimisation (cf. Scheme 1). In CNS, default number is 2000 (10 cycles of 200 iterations). However, in the combination of crystallographic refinement and QM calculations, divergence was observed unless only one iteration was used. However, this is not the case with ComQum-N (Table S9; it should be noted that the calculations in Tables S9–S11 used slightly different NMR restraint than in the other tables; therefore the NMR energies are larger). On the contrary, the convergence was faster with many iterations. Likewise, the total energy was lower. It turned out to be favourable to start the calculation with a normal NMR-restrained MM minimisation (without any QM) of the protein to convergence (~10 000 iterations), using the standard van der Waals parameters. Still, it can be seen that the results are not fully converged until 40 000 steps of MM minimisations are used. However, if that many steps are allowed, most of the time is spent in the MM minimisations and the full optimisations will take a very long time (several weeks). Therefore, we decided to use the default 2000 MM steps, for which the Ca–O distances are converged to within 0.01 Å and the total and NMR energies within 3 and 1 kJ/mol, respectively.

Third, we tested the effect of changing the weight of the NMR restraints. In ComQum-X, the results strongly depend on the relative weight between crystallography and the MM and QM energy functions (Ryde et al., 2002). For ComQum-N, the effect seems to be less pronounced: The average Ca–O bond length does not change at all (Table S10), whereas the individual distances change by up to 0.05 Å. Of course, the energy terms change more when increasing the NOE weight from 75 to 750. The NOE energy decreases by almost a factor of 4, whereas the other two terms increase slightly. However, the MM energy increases even more, so that the total energy increases by 446 kJ/mol. The energy of the quantum system changes by less than 10 kJ/mol (data not shown). Therefore, we see no reason to modify the NMR weights from their default values.

Finally, we also tested the treatment of electrostatics in the ComQum-N calculations. By default, CNS ignores all electrostatic interactions in NMR refinement. However, in the QM calculations, electrostatics within the QM system is included. We can then choose to include only these electrostatic interactions, include also the polarisation of the QM system by the surrounding protein (which is the standard choice in QM/MM optimisations), or even to turn on the electrostatics also in the NMR-restrained minimisation. We tried all three alternatives for several systems (Table S11 shows two typical cases). However, we always saw a strong increase of the NMR energies if electrostatics were included in NMR-restrained minimisation. This is in accordance with the consensus that NMR refinement should be run without electrostatics, unless the protein is explicitly solvated.

The other methods gave quite similar results with NMR energies within 10 kJ/mol. However, it was invariably observed that calculations without any point charges in the QM calculations gave a lower energy than with the point charges (by 4–7 kJ/mol). Furthermore, calculations where the QM system is dispersed into a continuum solvent (COSMO method, Klamt and Schüürmann, 1993) gave even lower NMR energies (by 1–5 kJ/mol). We doubt that these results are general, because the energies involved are so small and because the effect of point charges and COSMO should depend on the detailed structure of the surroundings. For water-exposed sites, as the present Ca^{2+} sites, it is possible that continuum calculations with a dielectric constant of ~80 may improve the results. However, for sites that are buried inside the protein and interacts with the surroundings with many hydrogen bonds, it is likely that a point-charge model would be the best choice.

For the general use of ComQum-N, our best recommendation is to run without any point charges in the QM calculation, unless there is extensive hydrogen bonding to the site of interest, and without any continuum model. This would be in accordance with the treatment of the surrounding protein. However, the optimum solution would probably be to include electrostatics in all calculations (i.e., both for the QM system and the surrounding protein). Then it would also be necessary to include full solvation of the protein by explicit water molecules. CNS is not set up and calibrated for such calculations, whereas other programs, e.g., AMBER (Case et al., 2002), allow for such an approach.

Refinements with the ComQum-N method

After this calibration of ComQum-N, we run production calculations on EGF34. As for standard NMR refinement, our aim is to obtain an ensemble of possible structures of the Ca^{2+} sites. Therefore, we started ComQum-N calculations for each of the two Ca^{2+} sites from the ten best structures (in terms of total energy) in Tables S6, S7, and S8, i.e., for both the mono- and bidentate sites. Of course, we could also have started from structures obtained with other methods, but for structures obtained without or with only O–O restraints, this would have been waste of computer resources, because the starting structures are too poor and there is essentially no force in QM to assemble the Ca^{2+} site if the ligands are far away.

For the same reason, we performed only the final minimisation of the Ca^{2+} site, i.e., no high-temperature dynamics was run. Such a local optimisation of the Ca^{2+} site in an ensemble of structures obtained by standard NMR refinement exploit the computer resources in the best way; if the ComQum-N method had been used already in the early phases of the refinement, most of the structures would have ended up with unrealistic Ca^{2+} sites (like those in Table S2) at a very high computational cost. Therefore, it is more favourable to produce reasonable starting structures for the Ca^{2+} sites with the MM-restraint methods and then perform a final minimisation of the Ca^{2+} site with ComQum-N. Provided that the weight factors are appropriate, this will still allow for significant modifications of the sites, if necessary.

The results show that in many of the final structures, the coordination has changed (Tables S12 and S13). In five of the calculations starting from the monodentate site, the final structure is bidentate (four with Asp and one with Glu). Moreover, in seven of the calculations (mostly for site 2), one of the carbonyl groups dissociate (to Ca–O distance of 3.04–3.67 Å; 4.46 Å in one case), but in two cases this is compensated by the bidentate binding of Asp. Likewise, only four of the bidentate Glu structures keep all the seven ligands, whereas four of them become monodentate, seven become six-coordinate with the Glu ligand still bidentate, two become six-coordinate with the Asp ligand bidentate, three become five-coordinate, and two actually become seven-coordinate with both the Asp and Glu ligand bidentate, but with the water ligand dissociated. Finally, seven of the bidentate Asp sites keep all the ligands, whereas two become monodentate, nine lose one ligand, and two lose two ligands (not Asp). Thus, in total, there are 14 monodentate structures, 5 bidentate structures with Glu, 11 bidentate structures with Asp, 19 bidentate structures with one lost ligand (12 with Asp and 7 with Glu), 9 five-coordinate structures, and 2 seven-coordinate structures with both Asp and Glu bidentate. Typical examples of the monodentate and bidentate structure with Asp are shown in Figure 3.

The average Ca–O distances follow the coordination number of the site: The five-coordinate sites have an average Ca–O distance of 2.38 Å, the six-coordinate sites have average Ca–O distances of 2.42–2.44 Å, and the seven-coordinate sites have average Ca–O distances of 2.49–2.51 Å (in all these averages, the dissociated ligands have

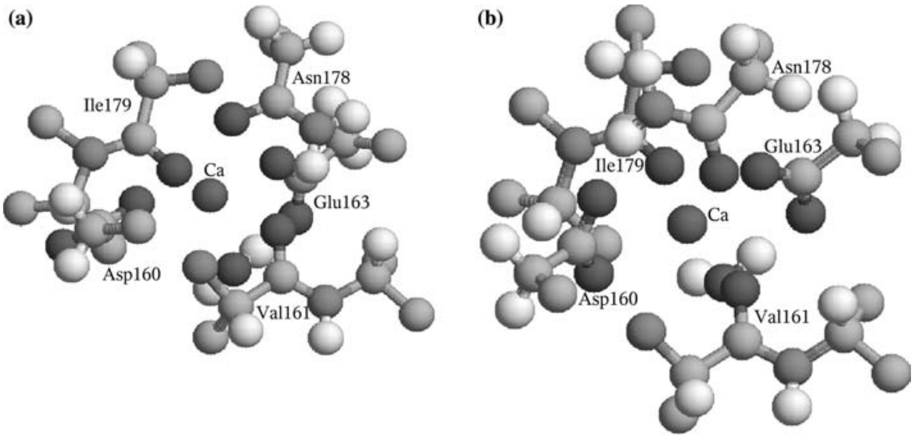

Figure 3. The final ComQum-N structures of Ca^{2+} site 1 with a monodentate (a) or bidentate with Asp (b) binding. The best structures (in terms of total energy) from Tables S12 and S13 were used.

been excluded, in variance to the averages in Tables S12 and S13). The Glu and Asp residues give the shortest Ca–O bonds (average 2.36 and 2.37 Å; the shortest bond encountered was 2.23 Å for Glu in a five-coordinate site). The Asn ligand gives slightly longer distances (2.39 Å), whereas water and the two carbonyl groups give the longest bonds (averages 2.49–2.57 Å). The longest bond encountered was 2.89 Å for the first carbonyl in a monodentate site 1 (our limit for a dissociated ligand was 3.04 Å).

Looking on the energies, it can be seen that the NMR energies are moderate, 98–188 kJ/mol (average 140 kJ/mol). This is slightly higher than for the structures obtained without any Ca restraints, but similar to all the calculations with restraints (averages of the 20 best structures of 121–155 kJ/mol). Thus, the change in van der Waals parameters and the heavy initial MM minimisation do not significantly affect the energies. It is hard to discern any clear trends among the various coordination modes, except that all sites with a bidentate Glu ligand have relatively high NMR energies (averages 151–171 kJ/mol compared to 129–140 kJ/mol for the other types of sites). The five-coordinate structures and six-coordinate structure with a bidentate Asp give a slightly lower average NMR energy (129 and 131 kJ/mol) than the monodentate and bidentate structures with Asp (137 and 140 kJ/mol). The lowest NMR energies are obtained for two six-coordinate structures with a bidentate Asp ligand, whereas the mono- and bidentate Asp sites give the lowest QM and total energies. There is a clear correlation between the NMR and total energies, whereas there are hardly any correlation between the QM energy and the other energies (possibly a slight anticorrelation for the dissociated sites). Interestingly, for all coordination modes, the QM energies are ~20 kJ/mol lower for site 1, whereas the NMR energies are ~10 kJ/mol lower for site 2. This most likely reflect that there are more NMR restraints for site 2 than for site 1.

In conclusion, the ComQum-N method works properly and give a general structure of the protein similar to the other methods, as can be seen in Figure 4 (note that there are three times as many structures in these than in Figure 1). Both the ComQum-N and MM results quite clearly show that a bidentate binding of the Glu ligand is energetically unfavourable and therefore unlikely. However, we cannot unambiguously decide if the Ca^{2+} sites are monodentate or bidentate with Asp. These two structures give similar energies (NMR, QM, and total) and they also arise in calculations started from other coordination modes. Perhaps, a slightly higher tendency to bidentate Asp coordination can be seen for site 2. It is even likely that the two types of sites may show a fast interchange on an NMR time-scale, because the barrier between the mono- and bidentate binding of a carboxylate group is small (e.g., ~10 kJ/mol for Zn^{2+} complexes, Ryde, 1999).

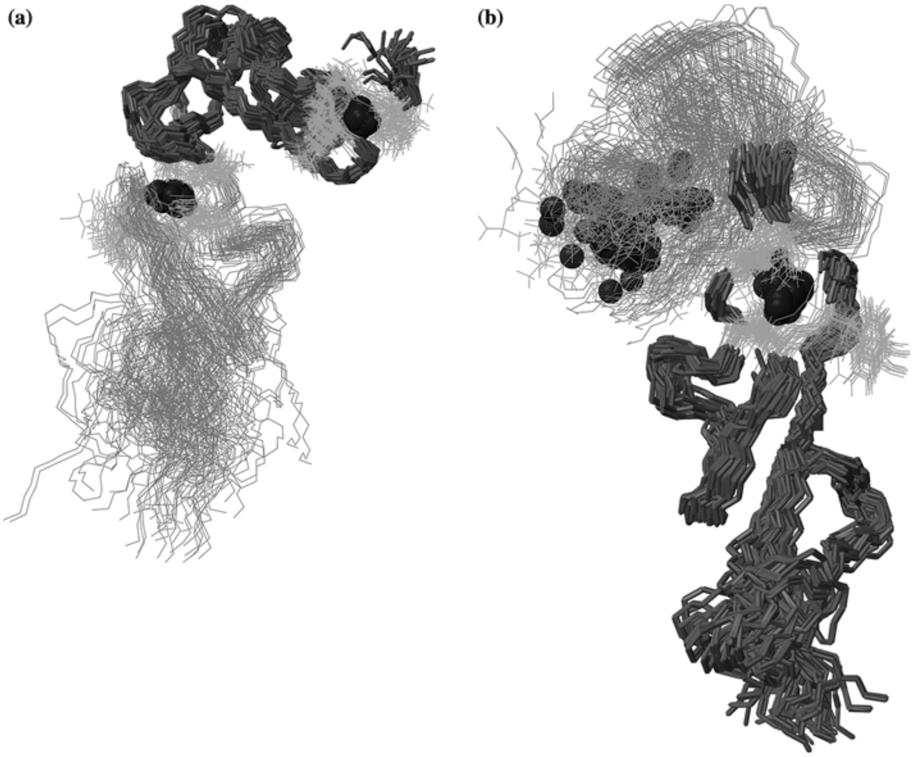

Figure 4. General structure of the EGF34 fragment with the ligands of the two Ca sites emphasized in green and with the Ca^{2+} ions in black. All 60 ComQum-N structures are superimposed, using the data in Tables S12 and S13. Note that this is three times as many structures as in Figure 1 (and Figures S1 and S2). The left- and right-hand side images were obtained by superimposing domain 3 and 4, respectively.

Concluding remarks

In this paper, we have tested and compared a number of methods to treat a metal site in NMR protein structure refinement. In particular, we have developed and tested two new methods to employ QM data in the refinement to supplement the refinement and obtain a more accurate description of a site of interest.

We have seen that it is not enough to describe the site as simple ligand–ligand restraint: This may lead to a structure in which the putative ligands do not have the proper orientation to bind the metal (Figure 2). Instead, explicit metal–ligand bonds seem to be necessary to yield a realistic metal site.

QM data can be introduced in the refinement either directly, as in the ComQum-N method or via the construction of an accurate MM potential. The latter method (Hess2FF) has been developed and tested out for hetero-compounds in crystal structures (Nilsson et al., 2003). However, it is equally suited and applicable for NMR refinement. The present results (Tables S6–S8) shows that it performs quite well also for the theoretically complicated plastic Ca^{2+} sites.

We have also developed the ComQum-N method, as an NMR variant of crystallographic quantum refinement (Ryde et al., 2002; Ryde and Nilsson, 2003b). It is an appreciably more accurate method, because it avoids the risk of introducing errors during the conversion of QM data to the MM potential. Moreover, ComQum-N allows changes in the coordination number during the refinement, reducing the risk of biasing the results by the choice of the MM potential. This is nicely illustrated in the application to EGF34, for which the unexpected possibility of a bidentate binding of the Asp ligand was discovered by the initial ComQum-N calculations. On the other hand, this also means that there is a risk that ComQum-N looses ligands by chance, e.g., if the starting structure is poor (as was seen in several of the ComQum-N calculations on EGF34). In this sense, ComQum-N is less robust than an MM potential, because there is only a minor attraction between the ion and its ligand at long distances. Therefore, ComQum-N cannot be used to construct the metal site during early phases of the refinement. Instead, a more robust method has to be used to construct the starting structure for the final local refinement with ComQum-N. On the other hand, ComQum-N allows the site to disrupt if it is not supported by the NMR restraints.

Another advantage with the MM method is of course the speed. With only an MM potential, the NMR refinement is as fast as standard refinement, meaning that an ensemble of 200 structures can be constructed within a few hours on a standard PC. However, the QM optimisation of the metal site and the frequency calculation takes appreciably longer time, typically several days. Sometimes, the QM system may be so large (over ~50 atoms) that it becomes hard to perform the frequency calculation. The ComQum-N calculations, on the other hand, typically take 1 or 2 days each, meaning that a full ensemble of 200 structures would probably take a prohibitively long time. Therefore, the ComQum-N calculations have to be restricted to the ~10 best structures obtained by other methods. However, ComQum-N involves only energy and force calculations and therefore avoid the frequency calculations, which have a much larger demand of memory and disk space. Therefore, ComQum-N can be run on appreciably larger systems than the frequency calculation (up to ~200 atoms).

Finally, it can be noted that additional experimental information can easily be included in the Hess2FF MM potential. For example, data from crystal structures (either proteins or small molecules) can be included. On the other hand, QM calculations using standard density functional methods typically give geometries of metal sites of an accuracy that is better than in a single protein crystal structure (Ryde and Nilsson, 2003a), so this is normally not advantageous except when accurate small-molecule crystallographic data are present for exactly the metal site of interest (the type of metal ligands strongly affects the bond lengths also of the other ligands to the metal) (Nilsson et al., 2003).

In conclusion, we suggest the following approach for the treatment of metal sites in NMR structure refinement: If the main interest is the general fold of the protein, a simple MM potential with only metal–ligand bonds is probably the best way to model the metal site. In this case, QM calculations are not necessary; instead the ideal bond length can be estimated from crystal structures of similar sites or from chemical intuition, and dummy force constants (~100 kJ/mol/Å^2) can be employed. However, if a more detailed picture of the metal site is intended, a better method is needed. Hess2FF is recommended when a large number of different systems is to be tested, whereas ComQum-N is the best method when accurate results are needed, e.g., at the end of an investigation with Hess2FF. The ComQum-N calculations can be started from a refinement with a simple bonded potential.

The present QM calculations have been performed with density functional theory and medium-sized (6-31G*) basis sets. We think this is an appropriate level of theory for the general use of our methods, even if it is quite costly (a few days of CPU time). However, for simpler systems (e.g., normal organic molecules), a lower level of theory could be used, e.g., the semiempirical PM3 method (Stewart, 1989), or even accurate MM methods, such as MMFF (Halgren, 1996), could be used (Nilsson et al., 2003). Such calculations can be performed within an hour for most systems of interest.

It is important to note that the presented methods, Hess2FF and ComQum-N, are not restricted to metal sites. On the contrary, they are fully general and can be used to any site of interest. However, for the normal amino acids and nucleic acids, quite accurate target values for bond lengths and angles exist (Engh and Huber, 1991), reducing the need of more accurate methods. Yet, for unusual molecules (hetero-compounds), such as substrates and inhibitors, no MM potentials exist and for such sites, the present methods could be used. In particular, they could be useful for high-throughput NMR structure determination, where automatic methods are needed for hetero-compounds.

An important use of the present methods is to test conflicting structural hypotheses, such as whether a ligand binds in a mono- or bidentate mode in the present investigation. This is done by refining both candidates and comparing their energies and structures. By similar methods, it has been possible to decide the protonation state of metal-bound solvent molecules by ComQum-X (Ryde and Nilsson, 2003b; Nilsson and Ryde, 2004).

Another possible application of ComQum-N is for structures of proteins that contain paramagnetic metal ions. Such metals lead to a significant broadening of the NMR signals around the metal site so that the local structure is hard to determine (Banci et al., 2002; Arnesano et al., 2003). By the use of ComQum-N, accurate information about the missing local structure around the metal ion could be obtained. Finally, ComQum-N can provide ideal starting structures for QM investigations of the structure, function, and reaction mechanism of proteins for which only NMR structures are available, providing an optimum compromise between experiments and quantum chemistry.

Acknowledgements

This investigation has been supported by grants from the Swedish research council (VR) and by computer resources of Lunarc at Lund University.

Supplementary material to this paper is available in electronic form at http://dx.doi.org/10.1007/s10858-004-6729-7.

References

Ahlrichs, R., Bär, M., Baron, H.-P., Bauernschmitt, R., Böcker, S., Ehrig, M., Eichkorn, K., Elliott, S., Haase, F., Häser, M., Horn, H., Huber, C., Kölmel, C., Kollwitz, M., Ochsenfeld, C., Öhm, H., Schäfer, A., Schneider, U., Treutler, O., von Arnim, M., Weigend, F., Weis, P. and Weiss, H. (2000) *TURBOMOLE* Version 5.6, Universität Karlsruhe, Germany.

Appella, E., Weber, I. and Blasi, F. (1988) *FEBS Lett.*, **231**, 1–4.

Arnesano, F., Banci, L., Bertini, I., Felli, I.C., Luchinat, C. and Thompsett, A.R. (2003) *J. Am. Chem. Soc.*, **125**, 7200–7208.

Banci, L., Pierattelli, R. and Villa, A.J. (2002) *Adv. Protein Chem.*, **60**, 397–449.

Becke, A. (1988) *Phys. Rev. A*, **38**, 3098–3100.

Brunger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Niges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (2000) *Crystallography & NMR System CNS*, Version 1.1. Yale University.

Campbell, I.D. and Bork, P. (1993) *Curr. Opin. Struct. Biol.*, **3**, 385–392.

Case, D.A., Pearlman, D.A., Caldwell, III, J.W., T.E.C., Wang, J., Ross, W,S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V., Cheng, A.L., Vincent, J.J., Crowley, M., Tsui, V., Gohlke, H., Radmer, R.J., Duan, Y., Piteral, J., Massova, I., Seibel, G.L., Singh, U.C., Weiner, P.K. and Kollman, P.A. (2002) *Amber Version 7*, University of California, San Francisco, U.S.A.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G. and Skelton, N.J. (1996) *Protein NMR Spectroscopy. Principles and Practice*, Academic Press, London.

Cruickshank, D.W.J. (1999) *Acta Crystallogr.* **D55**, 583–601.

de Silva, J.J.R.F. and Williams, R.J.P. (1991) *In the Biological Chemistry of the Elements*, Clarondon, Oxford.

Doreleijers, J.F., Raves, M.L., Rullmana, T. and Kaptein, R. (1999) *J. Biomol., NMR*, **14**, 123–132.

Downing, A.K., Knott, V., Werner, J.M., Cardy, C.M., Campbell, I.D. and Handford, P.A. (1996) *Cell*, **85**, 597–605.

Eichkorn, K., Treutler, O., Öhm, H., Häser, M. and Ahlrichs, R. (1995) *Chem. Phys. Lett.*, **240**, 283–290.

Eichkorn, K., Weigend, F., Treutler, O. and Ahlrichs, R. (1997) *Theor. Chim. Acta*, **97**, 119–124.

Engh, R.A. and Huber, R. (1991) *Acta Crystallogr.*, **A47**, 392–400.

Fields, B.A., Bartsch, H.H., Bartunik, H.D., Cordes, F., Guss, J.M. and Freeman, H.C. (1994) *Acta Crystallogr.*, **D50**, 709–730.

Halgren, T.A. (1996) *J. Comput. Chem.*, **17**, 490–641.

Hehre, W.J., Radom, L., Schleyer, P.v.R. and Pople, J.A. (1986) *Ab Initio Molecular Orbital Theory*, Wiley-Interscience, New York.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K. and Hood, L. (1997) *Science*, **278**, 609–614.

Hertwig, R.H. and Koch, W. (1997) *Chem. Phys. Lett.*, **268**, 345–351.

Jalilehvand, F., Spånberg, D., Lindqvist-Reis, P., Hermansson, K., Persson, I. and Sandström, M. (2001) *J. Am. Chem. Soc.*, **123**, 431–441.

Jensen, F. (1999) *Introduction to Computational Chemistry*, John Wiley & Sons, Chichester.

Klamt, A. and Schüürmann, J. (1993) *J. Chem. Soc., Perkin Transact. II*, **5**, 799–805.

Klamt, A., Jonas, V., Bürger, T. and Lohrenz, J.C.W. (1998) *J. Phys. Chem. A*, **102**, 5074–5085.

Kleywegt, G.J. and Jones, T.A. (1995) *Structure*, **3**, 535–540.

Kleywegt, G.J. and Jones, T.A. (1998) *Acta Crystallogr.*, **D54**, 1119–1131.

Maseras, F. and Morokuma, K. (1995) *J. Comput. Chem.*, **16**, 1170–1179.

Monard, G. and Merz, K.M. (1999) *Acc. Chem. Res.*, **32**, 904–911.

Mulholland, A.J. (2001) In *Theoretical Biochemistry – Processes and Properties of Biological Systems*, Vol. 9: *Theoretical and Computational Chemistry*, Eriksson, L.A. (Ed.), Elsevier Science, Amsterdam, pp. 597–505.

Nicoll, R.M., Hindle, S.A., MacKenzie, G., Hillier, I.H. and Burton, N.A. (2001) *Theor. Chim. Acta*, **106**, 105–112.

Nilsson, K. and Ryde, U. (2004) *J. Inorg. Biochem.*, **98**, 1539–1546.

Nilsson, K., Lecerof, D., Sigfridsson, E. and Ryde, U. (2003) *Acta Crystallogr.*, **D59**, 274–289.

Norrby P.O. and Liljefors, T. (1998) *J. Comput. Chem.*, **19**, 1146–1166.

Olsson, M.H.M. and Ryde, U. (2001) *J. Am. Chem. Soc.*, **123**, 7866–7876.

Perdew, J.P. (1986) *Phys. Rev. B* **33**, 8822–8824.

Rao, Z., Handford, P., Mayhew, M., Knott, V., Brownlee, G.G. and Stuart, D. (1995) *Cell*, **82** 131–141.

Ryde, U. (1996) *J. Comput. Aided Mol. Des.*, **10**, 153–164.

Ryde, U. (1999) *Biophys. J.*, **77**, 2777–2787.

Ryde, U. (2003) *Curr. Opin. Chem. Biol.*, **7**, 136–142.

Ryde, U. and Nilsson, K. (2003a) *J. Am. Chem. Soc*, **125**, 14232–14233.

Ryde, U. and Nilsson, K. (2003b) *J. Mol. Struc. (Theochem)*, **632**, 259–275.

Ryde, U. and Olsson, M.H.M. (2001) *Int. J. Quantum Chem.*, **81**, 335–347.

Ryde, U., Olsen, L. and Nilsson, K. (2002) *J. Comput. Chem.*, **23**, 1058–1070.

Ryde, U., Olsson, M.H.M., Roos, B.O. and Carlos, A.B. (2001) *Theor. Chum. Acta,* **105**, 452–462.

Saha, S., Boyd, J., Werner, J.M., Knott, V., Handford, P.A., Campbell, I.D. and Downing, A.K. (2001) *Structure*, **9**, 451–456.

Schäfer, A., Klamt, A., Sattel, D., Lohrenz, J.C.W. and Eckert, F. (2000) *Phys. Chem. Chem. Phys.*, **2**, 2187–2193.

Siegbahn, P.E.M. and Blomberg, M.R.A. (2000) *Chem. Rev.*, **100,** 421–437.

Sigfridsson, E., Olsson, M.H.M. and Ryde, U. (2001) *J. Phys. Chem. B*, **105**, 5546–5552.

Stenberg, Y., Linse, S., Drakenberg, T. and Stenflo, J. (1997a) *J. Biol. Chem.*, **272**, 23255–23260.

Stenberg, Y., Muranyi, A., Steen, C., Thulin, E., Drakenberg, T. and Stenflo, J. (1997b) *J. Mol. Biol.*, **293**, 653–665.

Stenflo, J., Stenberg, Y. and Muranyi, A. (2000) *Biochim. Biophys. Acta,* **1477**, 51–63.

Stewart, J.J.P. (1989) *J. Comput. Chem.*, **10**, 209–220.

Svensson, M., Humbel, S., Froese, R.D.J., Matsubara, T., Sieber, S. and Morokuma, K. (1996) *J. Phys. Chem.*, **100**, 19357.

Torrent, M., Musaev, D.G. and Morokuma, K. (2001) *J. Phys. Chem. B*, **105**, 322–327.

Tossavainen, H., Permi, P., Annila, A., Kilpeläinen and Drakenberg, T. (2003) *Eur. J. Biochem.*, **270**, 2505–2512.

Wang, X., Li, M.X., Spyracopoulos, L., Beier, N., Chandra, M., Solaro R.J. and Syske, B.D. (2001) *J. Biol. Chem.*, **276**, 25456–25466.