Do Quantum Mechanical Energies Calculated for Small Models of Protein-Active Sites Converge?[†]

LiHong Hu, Jenny Eliasson, Jimmy Heimdal, and Ulf Ryde*

Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, SE-221 00 Lund, Sweden

Received: March 31, 2009; Revised Manuscript Received: September 1, 2009

A common approach for the computational modeling of enzyme reactions is to study a rather small model of the active site (20−200 atoms) with quantum mechanical (QM) methods, modeling the rest of the surroundings by a featureless continuum with a dielectric constant of ∼4. In this paper, we discuss how the residues included in the QM model should be selected and how many residues need to be included before reaction energies converge. As a test case, we use a proton-transfer reaction between a first-sphere cysteine ligand and a second-sphere histidine group in the active site of [Ni,Fe] hydrogenase. We show that it is not a good approach to add groups according to their distance to the active site. A better approach is to add groups according to their contributions to the QM/MM energy difference. However, the energies can still vary by up to 50 kJ/mol for QM systems of sizes up to 230 atoms. In fact, the QM-only approach is based on the hope that a large number of sizable contributions will cancel. Interactions with neutral groups are, in general, short-ranged, with net energy contributions of less than 4 kJ/mol at distances above 5 Å from the active site. Interactions with charged groups are much more long-ranged, and interactions with buried charges 20 Å from the active site can still contribute by 5 kJ/mol to the reaction energy. Thus, to accurately model the influence of the surroundings on enzyme reaction energies, a detailed and unbiased atomistic account of the surroundings needs to be included.

Introduction

During the latest 20 years, theoretical calculations have been established as an important complement to experimental studies of enzyme reactions. In particular, they have been shown powerful to study details of enzyme mechanisms, giving structures and energies of putative transition states and reaction intermediates.[1−6]

Theoretical studies of enzyme reactions are a rather young science. Therefore, there is still no consensus how such calculations are ideally performed. On the contrary, two rather different approaches have been developed. In the first, which we will call the QM-only approach,[1−4] only a rather small model of the active site is explicitly studied. A typical size is 20−200 atoms, which corresponds to the substrate and up to 20 amino acids from the surroundings. The rest of the enzyme is either ignored or treated as a featureless continuum, characterized by a dielectric constant of ∼4. Owing to the small size of the studied system, accurate quantum mechanical (QM) methods can be used, and many different mechanisms may be compared. In addition, the studied system is so small that it is possible to have a full control of the conformations of all included groups (to ensure that all reaction intermediates reside in the same local minimum). Entropy effects are typically obtained by a normal-mode analysis of calculated harmonic frequencies. It is customary to perform geometry optimizations with a medium-size basis set and then calculate single-point energies with a larger basis set. Typically, it is necessary to fix the position of one or a few atoms in each residue to keep the structure reasonably close to the protein structure. This approach originates from QM calculations on simple organic reactions, where it has been applied with great success. The accuracy of the QM-only approach for enzymes is typically estimated to be 12−20 kJ/mol.[1,4]

In the alternative approach, the full enzyme is included by combined QM/MM calculations,[7] in which the active site is treated by QM methods whereas the rest of the enzyme and often some shells of surrounding water molecules are treated at the molecular mechanics (MM) level.[5,6] The size of the QM system is similar to that in the QM-only approach. The advantage with the QM/MM approach is of course that a detailed account of the surrounding protein is explicitly included. Thereby, the risk that the results are biased by the selection of residues to include in the model system is avoided. The disadvantage with this approach is the size of the simulated system. For a full protein with water solvation (typically 10 000−100 000 atoms), there is practically an infinite number of possible conformations. It is virtually impossible to ensure that the most relevant conformation is obtained, and there is a great risk that different states during the reaction may reside in different local minima. Unfortunately, the energy is sensitive to the conformation (a single extra hydrogen bond, even far from the active site, will change the energy of the system by ∼20 kJ/mol). Likewise, there is an uncertainty of the exact location of all atoms in the system because hydrogen atoms are normally not discerned in protein crystal structures (the typical starting point of QM/MM calculations). In particular, the number of protons (i.e., the net charge of the protein) is unknown. There are ways to deal with these problems, for example, by using ensembles of structures and free-energy methods, but they are quite time-consuming.[6,8−10] Thus, the QM/MM approach is clearly more complicated than the QM-only approach.

[†] Part of the "Walter Thiel Festschrift".

* To whom correspondence should be addressed. E-mail: Ulf.Ryde@teokem.lu.se. Tel: +46 - 46 2224502. Fax: +46 - 46 2224543.

10.1021/jp9029024 CCC: $40.75
Published on Web 09/29/2009

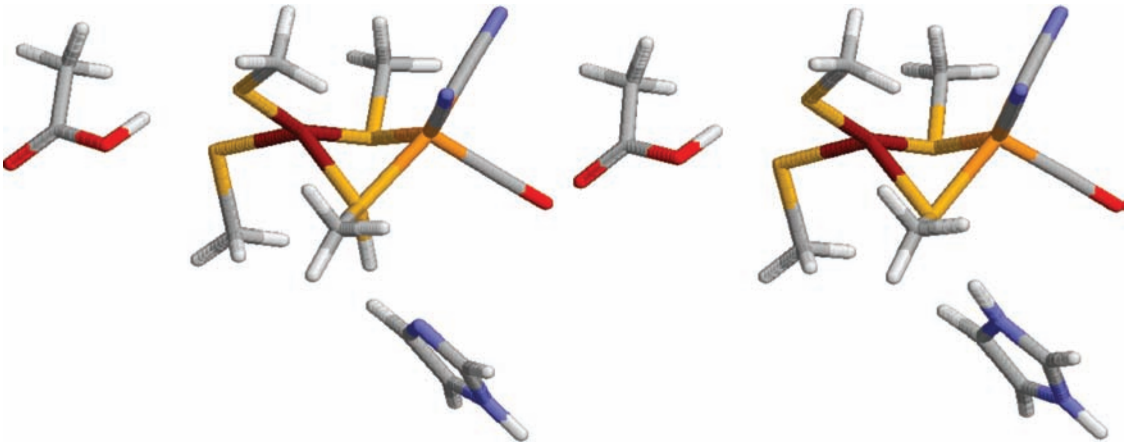

Figure 1. The studied reaction, with the HID state to the left and the HIP state to the right, illustrated with the smallest, 46 atom QM system without any added residues.

Apparently, both the QM-only and QM/MM approaches have their advantages and disadvantages, and currently, no consensus has been reached of which approach is preferable. In the present paper, we study the convergence of energies from the QM-only approach with respect to the size of the QM system. In particular, we test whether the results of QM/MM calculations can be used to guide the selection of the QM system in a QM-only approach, so that the advantages of the two methods can be combined.

Methods Section

In this paper, we study the reaction energy of the simple proton-transfer reaction in Figure 1. It involves the transfer of a proton from the S^{γ} atom of one of two cysteine ligands that bridge Ni and Fe ions in the active site of [Ni,Fe] hydrogenase (Cys-546 in *Desulfovibrio fructosovorans*) to the $N^{\varepsilon 2}$ atom of a second-sphere histidine ligand (His-79). We have calculated the energy difference between the form in which the proton resides on Cys-546 (called the HID state) and the form in which the proton resides on His-79 (called the HIP state).

We test if we can obtain a good estimate of this reaction energy with a rather small QM system, studied in vacuum or a continuum solvent with a dielectric constant of $\varepsilon = 4$.[1-3,5] We use the same geometries for all calculations in order to minimize the effects of the structure. The structure comes from a QM/MM minimization of the protein in the two states,[11] based on a 1.81 Å crystal structure.[12] The geometry of the surrounding protein is exactly the same in the two states outside of the 46 atom QM system. Thereby, we avoid the risk that the surroundings may reside in different local minima for the two states.[6]

We systematically enlarge the QM system up to the size when the calculations start to be prohibitively slow (450−700 atoms). Two different approaches are followed. In the first (Dist), we include residues in the order of their closeness to the active site (the shortest distance between any atom in the residue and any atom in the smallest QM model of the active site). In the second approach (Ene), we instead include them according to their contribution to the free energy,[11] according to a QM/MM free-energy perturbation approach, called QM/MM thermodynamic cycle perturbation (QTCP).[10] In addition, two different sizes of the added residues are tested. In the first (Res), the full residue is added, including capping CH_3CO- and $-NHCH_3$ groups from the previous and following residues. In the second (Grp), we instead add only functional parts of the amino acid, namely, either the backbone $CH_3CONHCH_3$ group (named according to the residue of the NH group), the side chain (the functional part of the polar groups, including a CH_3- buffer, except for the aromatic amino acids, for which the ring systems are truncated by a hydrogen atom), or the rest (which includes the whole side chain of the nonpolar and nonaromatic amino acids). These two schemes gives rise to four series of calculations, which will be called Dist-Res, Dist-Grp, Ene-Res, and Ene-Grp, respectively.

For each series, residues or groups were added systematically to the QM system, which initially consisted of Fe^{II} and Ni^{II} (both in the closed-shell low-spin state), their first-sphere ligands Cys-72, -75, -543, and -546 (modeled by CH_3S^-), CO, and two CN^- ligands, as well as the two second-sphere ligands His-79 (the proton acceptor, modeled by an imidazole group) and Glu-25 (modeled as a protonated acetic acid group because it is hydrogen-bonded to the S^{γ} atom of Cys-543). This system consists of 46 atoms and is shown in Figure 1. This was the QM system optimized by QM/MM for the starting structure. All broken bonds were truncated by a H atom at a C−H distance of 1.101 Å.

Then, single-point energies were calculated with the Becke 1988−Perdew1986 density functional[13,14] and the def2-SV(P) basis sets for all atoms.[15] The calculations were sped up by expanding the Coulomb interactions in auxiliary basis sets, the resolution-of-identity approximation.[16,17] Finally, all atoms were immersed into a continuum solvent with a dielectric constant of $\varepsilon = 4$, and the energy was recalculated using the continuum conductor-like screening model (COSMO).[18,19] These calculations were performed at the same level of theory and with default values for all parameters (implying a water-like probe molecule). For the generation of the cavity, a set of atomic radii has to be defined. We used the optimized COSMO radii in Turbomole (1.30, 2.00, 1.83, and 1.72 Å for H, C, N, and O, respectively) and 2.0 Å for the metals.[20]

Result and Discussion

We have studied a simple proton-transfer reaction in the active site of [Ni,Fe] hydrogenase (shown in Figure 1), namely, the transfer of a proton from the S^{γ} atom of one of the two cysteine ligands that bridge Ni and Fe ions (Cys-546 in *Desulfovibrio fructosovorans*) to the $N^{\varepsilon 2}$ atom of a second-sphere histidine ligand (His-79). We have calculated the energy difference between the form in which the proton resides on Cys-546 (called the HID state) and the form in which the proton resides on His-79 (called the HIP state). This reaction has been previously studied with several different theoretical methods and is known

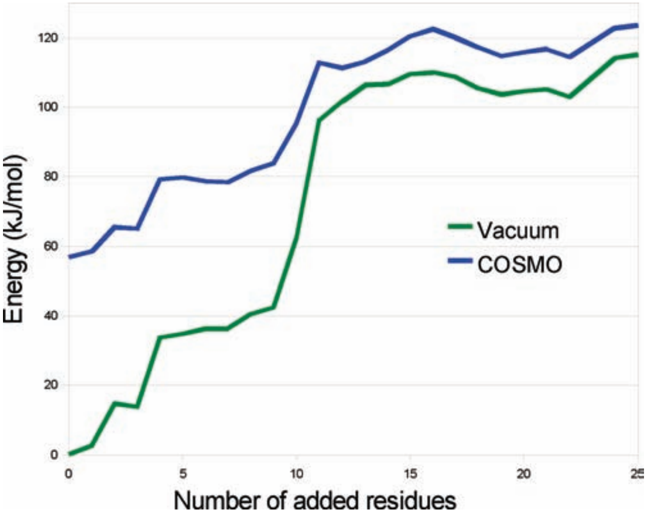

Figure 2. The results of the Dist-Res calculations.

to be sensitive to the treatment of the surrounding protein.[11,21] With a QM/MM free-energy perturbation approach (QTCP), using QM systems of 46−104 atoms, the HIP state is most stable by 74 kJ/mol (and 72 kJ/mol at a level of theory similar to the one used in this paper).

We have systematically (residue- or group-wise, as explained in the Methods Section) enlarged the QM system and calculated single-point energies in vacuum and in a continuum solvent with a dielectric constant of $\varepsilon = 4$ (called COSMO below). The residues or groups were added either according to their distance to the QM system (Dist) or according to their approximate QTCP energy components (Ene). This gave rise to four energy curves for the relative energy of the HID and HIP states, which will be discussed below.

The results of the Dist-Res calculations are shown in Figure 2. It can be seen that the vacuum energies vary between 0 and 115 kJ/mol (a positive sign indicates that HIP is the most stable state). The COSMO calculations are more stable and vary between 57 and 124 kJ/mol. However, once the 12 closest residues (up to 2.1 Å or 281 atoms) are included in the calculations, the energies stabilize at around 115 (vacuum) and 124 (COSMO) kJ/mol, with a variation of less than 13 kJ/mol. The largest contributions (shown in Table 1) come from three residues, Arg-476, which forms a hydrogen bond to one of the CN^- ligands (cf. Figure 3; 18 kJ/mol with COSMO), the backbone of Cys-546 (the proton donor, 14 kJ/mol), and His-481, which forms a hydrogen bond to the other N atom of the proton acceptor His-79 (11 kJ/mol). However, it is somewhat alarming that the 24th residue, a water molecule at 2.7 Å distance, still has an effect of 8 kJ/mol. After 25 residues have been added, the size of the QM system is 443 atoms.

Figure 4 shows the corresponding results of the Ene-Res calculations. It can be seen that the energies are actually more stable than those for the Dist-Res calculations, with a variation of −15 to +71 kJ/mol in vacuum and 49−94 kJ/mol with COSMO. Again, the results are reasonably stable, once the eighth residue has been added (233 atoms), around 62 and 84 kJ/mol in vacuum and with COSMO, respectively, and with a variation of less than 32 kJ/mol in vacuum and 24 kJ/mol with COSMO. The residues that give the largest contributions (Table 1) are Arg-476, Cys-546, and His-481 (18, 17, and 12 kJ/mol), as before, but also Thr-68, which is 3.2 Å from the proton acceptor His-79 (−21 kJ/mol), Arg-70, which is 4.5 Å from Cys-72 (17 kJ/mol), and Asp-541, which forms hydrogen bonds to Arg-476 (−12 kJ/mol; cf. Figure 3). Most of the residues added are either connected to the active-site residues or have a net charge (but all of the added charged residues are buried within the protein). When 25 residues have been added, the QM system contains 695 atoms, and each calculation takes about two weeks.

Apparently, it is quite uneconomical to include full residues in the calculation, leading to a practical limit of ∼25 residues. Therefore, we developed the group-wise method also and applied it to both schemes of addition. The results of the Dist-Grp calculations are shown in Figure 5. It can be seen that the results are quite similar to those in Figure 2. In particular, the energies stabilize at around 115 and 124 kJ/mol for the vacuum and COSMO calculations, respectively, after the addition of 18 groups (2.2 Å or 217 atoms), with a variation of less than 16 (COSMO) or 25 kJ/mol (vacuum). Again, the most important residues (Table 1) are Arg-476 (18 kJ/mol), the backbone of Cys-546 (14 kJ/mol), and His-481 (11 kJ/mol). In addition, Asp-114 at 3.3 Å distance, which also forms hydrogen bonds to Arg-476 (Figure 3), has a contribution of −10 kJ/mol. When 40 groups have been added, there are 396 atoms in the QM system.

Finally, Figure 6 shows the results of the Ene-Grp calculations. The results are quite similar to those of the Ene-Res calculations. The COSMO calculations are fairly constant at around 62 kJ/mol throughout the whole series, with variations of up to 31 kJ/mol. This variation is not reduced below 20 kJ/mol until 21 residues have been added (271 atoms). The largest contributions (Table 1) come from the side chains of Arg-476 (15 kJ/mol), Asp-541 (−13 kJ/mol), His-481 (11 kJ/mol), and Glu-S22 (from the small subunit, −11 kJ/mol), which is 5.6 Å from the QM system, as well as from the amide group between Ala-545 and Cys-546 (10 kJ/mol). After 40 groups have been added (including all groups that give QTCP components larger than 4.4 kJ/mol), there are 446 atoms in the QM system, as shown in Figure 3. It can be seen that the residues are preferentially located along the reactive $S^{\gamma}-H-N^{\varepsilon 2}$ reaction coordinate, illustrating that the observed effects come mainly from electrostatic interactions with a difference in the dipole moment of the active site, oriented along this reaction coordinate. Moreover, negatively charged residues on the left-hand side of Figure 3 give a negative contribution and positively charged residues a positive contribution, whereas on the right-hand side, the opposite is true (Asp-60 is the only exception), indicating the sign of the difference in dipole moments.

Comparing all four approaches, it is very alarming that the Dist and Ene approaches do not converge to the same results. Even after 40 groups are included, the results of the two approaches differ by over 60 kJ/mol (124 kJ/mol for Dist-Grp but 62 kJ/mol for Ene-Grp). The major reason for the difference is that the Dist approach does not include all energetically important residues (residues included in Ene-Grp but not in the Dist-Grp contribute by −35 kJ/mol to the energy). However, there is also a smaller contribution from residues included in the Dist-Grp but not in the Ene-Grp, −10 kJ/mol, and a similar contribution from nonadditive effects (i.e., from the fact that the groups are polarized differently in the two calculations, −15 kJ/mol). Thus, neither of the approaches is highly accurate, but since the former contribution is largest, this approach seems most reasonable.

However, it has recently been show that once all residues close to the active site have been included in QM-only calculations,[22,23] that is, something similar to the Dist-Grp approach, the calculated energy differences no longer depend on the selection of the dielectric constant in the continuum solvent model. From Figures 2 and 5, it can be seen that our results confirm this. Once ∼13 groups have been added

TABLE 1: Energy Contributions of the Most Important Residues or Groups in the Various Calculations (kJ/mol)[a]

	Dist-Res	Ene-Res	Dist-Grp	Ene-Grp		Dist+Ene		
residue	COSMO	COSMO	COSMO	COSMO	vacuum	COSMO	QTCP	distance (Å)
Arg-476	17.5	18.3	17.6	15.4	30.1		24.4	2.2
Asp-541		−11.7		−13.2	−21.2	−12.6	−16.0	4.8
Glu-S22		−3.8		−11.1	−9.3	−14.4	−9.8	5.6
His-481	11.3	12.0	10.6	10.7	15.4		9.5	2.0
Cys-546[b]	14.1	17.4	14.1	10.1	12.0		22.8	1.7
Asp-114		−12.2	−9.6	−8.6	−19.7	−10.2	−16.9	3.3
Asp-126				−7.3	−6.3	−0.8	−5.0	14.0
Glu-S75				−7.2	−4.8	−11.1	−5.8	7.0
Arg-23				7.0	5.3	5.2	5.1	11.9
Gly-547[b]				6.9	11.7		−4.6	1.7
Ala-80[b]				−6.5	−2.0		−6.2	2.9
Arg-428		−3.8		6.0	7.7	4.2	8.2	7.9
Gln-69[b]			5.6	5.7	6.3	4.9	13.6	3.2
Mg-site[c]		−0.5		5.7	32.8		−8.4	4.2
His-79[b]	−1.1	−1.4	−1.1	5.3	12.9		−9.2	1.6
Cys-75[b]	7.1	5.1	7.1	5.2	4.3		16.2	1.6
Arg-85		−9.1		−5.0	0.1		−5.7	9.6
Wat	8.3	7.2	8.2	−4.9	−11.0		9.1	2.7
Cys-72[b]	1.6		1.6	−4.3	−5.7		15.6	1.6
Arg-70		17.5		−3.7	−1.5		−7.3	7.3
His-538		5.4		3.6	7.7		6.7	8.9
His-115		−8.8		2.9	7.4		6.9	6.4
Ile-544[b]				−2.8	−3.4		−30.7	1.8
Cys-543[b]	−0.5	−0.4	−0.5	0.2	1.3		10.2	1.6
Thr-68		−21.3	1.1				−8.6	3.2
Glu-25[b]	0.6		0.6				0.3	1.6

[a] All COSMO contributions larger than 5 kJ/mol are included. The distance to the original QM system is also indicated. Note that the contributions correspond to residues in the Res calculations but to groups in the other calculations. A positive sign of the energies indicates that the HIP state is favored. [b] Backbone rather than the side chain. [c] The Mg site includes the Mg^{2+} ion, three water molecules, as well as the side chains of Glu-53 and His-630, and the backbone O atom of Gln-540.

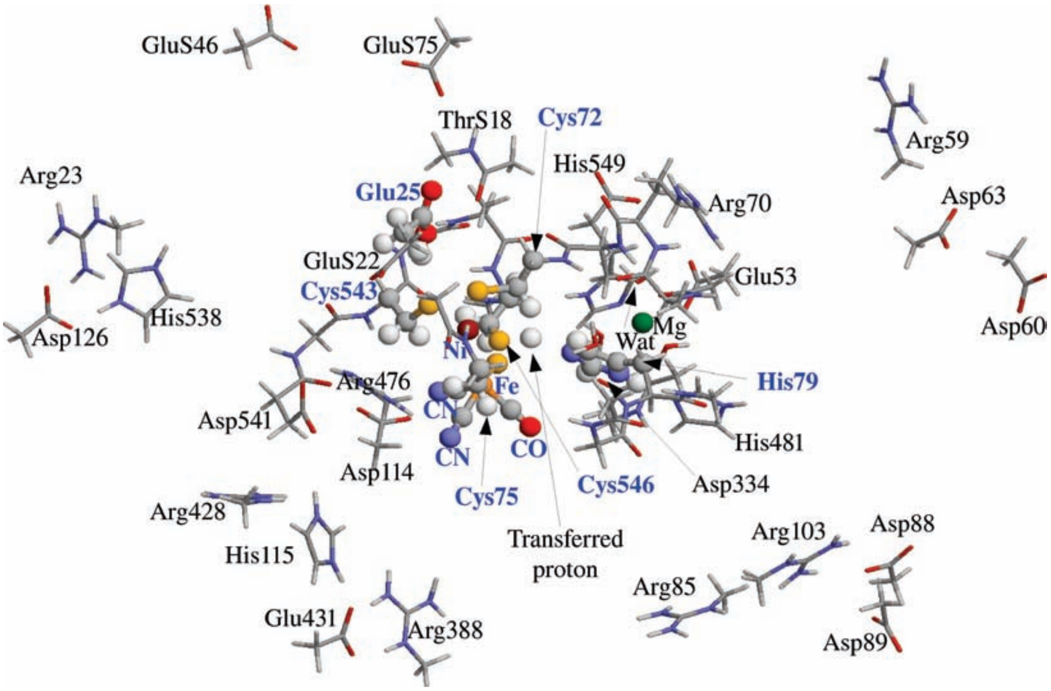

Figure 3. All groups included in the largest Ene-Grp calculation (446 atoms). The Mg ligand Gln540 is mainly hidden behind His79. Note that there are four water molecules in the calculation, one at the arrow, one just to left of the label, and two just to the right of the label. All are Mg ligands, except the upper one to the right of the label. The original quantum system is shown in balls and sticks and with blue boldface labels.

(220−280 atoms), the difference between the vacuum and COSMO calculations is less than 12 kJ/mol. Similar results are also observed for the Ene approach (Figures 4 and 6), but the convergence is slower, with differences of 11−16 kJ/mol after the addition of 35 groups (~400 atoms). This is of course an important result, showing that the general dielectric effect of the surroundings is saturated for systems of ~200 atoms with the Dist approach, making superfluous the use of the questionable continuum solvation model with its arbitrary dielectric constant.

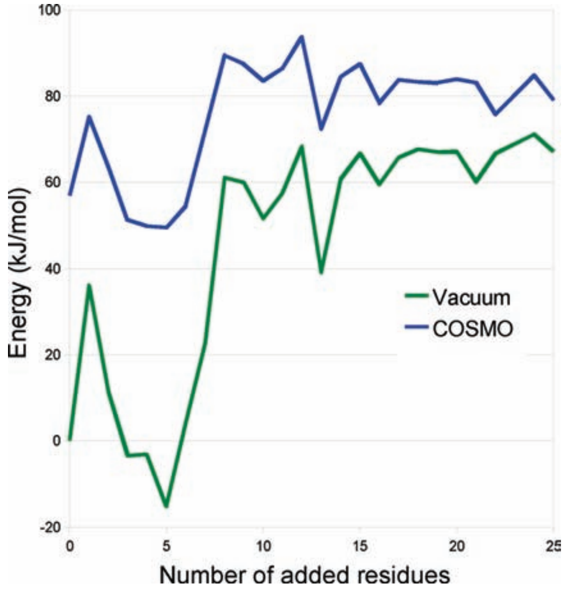

Figure 4. The results of the Ene-Res calculations.

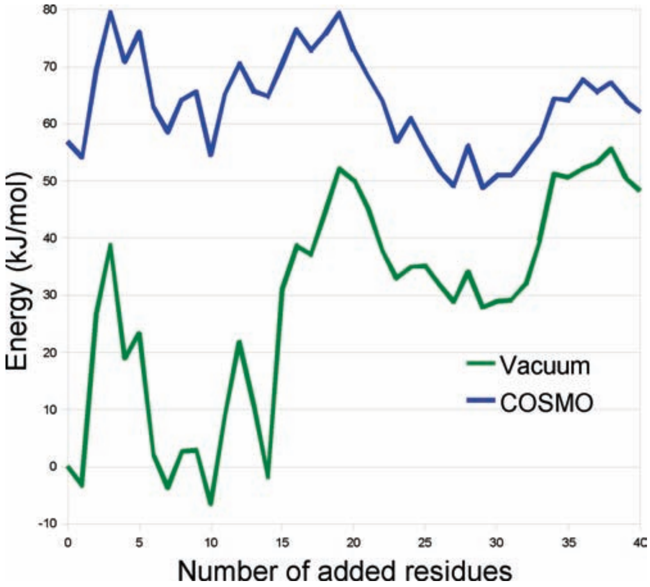

Figure 6. The results of the Ene-Grp calculations.

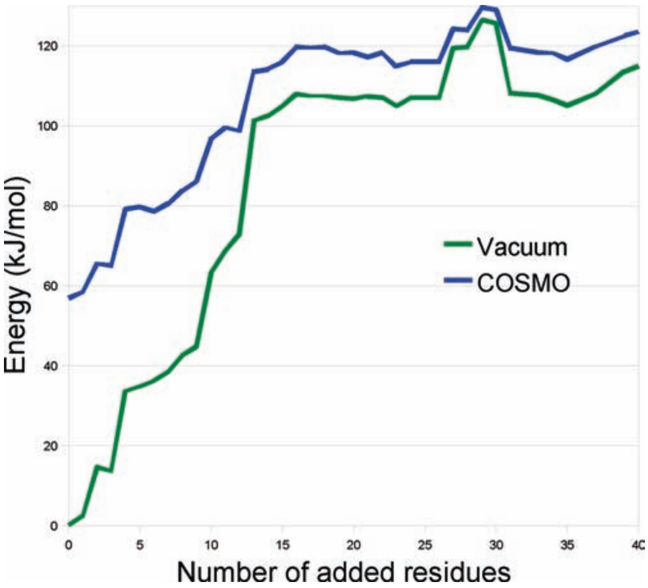

Figure 5. The results of the Dist-Grp calculations.

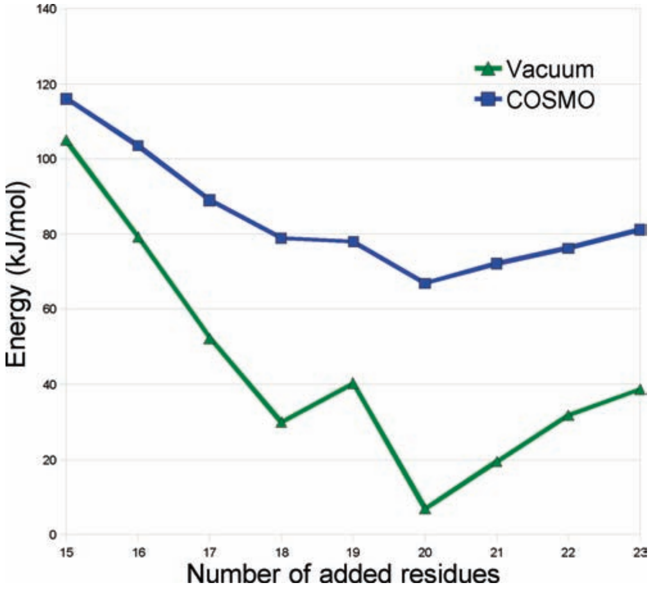

Figure 7. The results of the combined Dist+Ene approach.

Interestingly, this decreased dependence of the results on the dielectric constant has been taken as evidence of convergence of the results with respect to the number of atoms in the QM system.[22,23] However, this is not a valid conclusion, as is shown in Figure 7. There, we have started from the Dist-Grp result with 15 added residues, which, according to Figure 5, is essentially converged with respect to both the total energy and the effect of the COSMO model. However, if we now add the groups that gave the largest energy contribution according to the Ene-Grp approach (we call this approach Dist+Ene), the energy changes by up to 98 kJ/mol in vacuum and 49 kJ/mol with COSMO (after the addition of five residues), and the effect of the COSMO model increases to up to 60 kJ/mol. After the addition of the eight most important residues, the energy with COSMO is ~80 kJ/mol, that is, ~10 kJ/mol higher than the Ene-Grp result, as expected from the discussion above, but each added residue still contributes by ~5 kJ/mol. This shows that it is very hard to converge QM-only energies and that a small dependence on the dielectric constant does not imply that the size of the QM system is large enough. The reason for this is that the continuum solvent model can only take into account a featureless solvent, but it can never model specific strong interactions in the surrounding protein, such as charged groups.

Finally, we tested whether the interactions with charged groups can be screened by nearby groups. For the residue that gave the largest effect in Figure 7, Glu-S22 (the second added residue), we enhanced the model with all groups within 4 Å of the acetate model of Glu-S22, 109 atoms. However, this changed the COSMO energy by only 2 kJ/mol, that is, it reduced the effect of this residue from −14 to −12 kJ/mol. In vacuum, the effect was somewhat larger, from −27 to −19 kJ/mol. Thus, the effect of the charged residues cannot be decreased by adding nearby neutral groups. Therefore, such interactions need to be taken into account explicitly in a QM-only approach.

Next, we studied how to select groups to be included in the calculations. Figure 8 shows the relation between the estimated group contributions calculated at the QTCP and QM-only (Ene-Grp results with COSMO) levels. It can be seen that there is a fair correlation between the two estimates ($r^2 = 0.51$). However, the QM estimates are typically smaller than the QTCP estimates

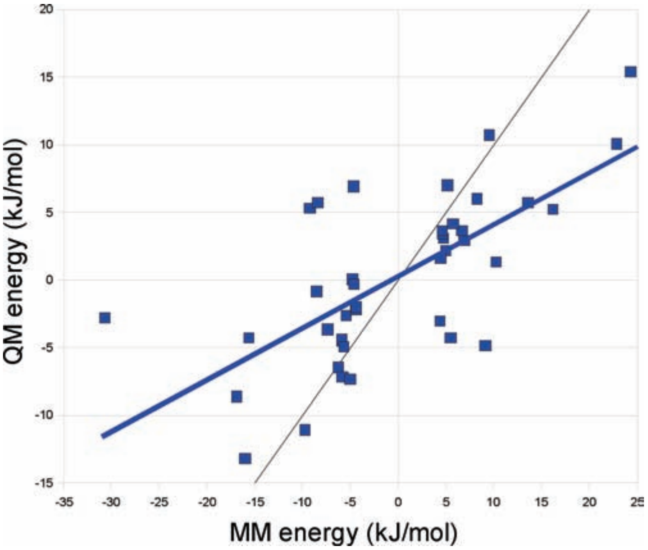

Figure 8. The relation between estimated QTCP and QM-only (Ene-Grp with COSMO) group contributions to the energy difference between the HIP and HID states.

(by a factor of 2.4 on average) as a consequence of the damping effect of the continuum solvation. On the other hand, there is no systematic shift in the two estimates (the intercept of the best correlation line is 0.2 kJ/mol), and for only six of the 40 groups do the two estimates give a different sign (and only for contributions smaller than 10 kJ/mol). Among the 10 and 20 largest QTCP energy contributions, 5 and 14 of the largest QM contributions are included, respectively. This shows that the QTCP components give a rough estimate of what groups are most important in the QM-only calculations, that is, that the Ene approach works fairly well to select the most important residues. It should be noted that both the QTCP and QM-only energy contributions are approximate, and it cannot be said which of them is more accurate. The QTCP contributions are free energies (and therefore not directly comparable to the QM pure energies), but they have been obtained with a linear approximation and without taking any account of the relaxation of the surrounding residues and solvent. Moreover, no polarization of the MM system is considered. On the other hand, the QM-only energies are obtained with a featureless continuum model of the surroundings, assuming that a dielectric constant of 4 is applicable to any part of the protein. In addition, only a few groups in the surroundings are considered explicitly. Owing to the many-body character of polarization, residue contributions are an ill-defined concept also in the QM-only approach. For example, even if there is a better correlation ($r^2 = 0.86$) between the residue contributions in the Ene-Grp and Dist+Ene approaches (where exactly the same groups are added but in a different order and to different QM systems), there are differences of up to 7 kJ/mol (Table 1).

Figure 9 shows the relation between the absolute energy contribution of each group and their distance to the QM system, marking interactions with charged groups by squares and those with neutral groups by diamonds. It can be seen that interactions with neutral groups are quite short-ranged; only groups within 4.5 Å of the QM system give significant contributions. However, it is possible that some important interactions are not included among the 40 added groups because the second to last neutral residue contributes by 6 kJ/mol (the backbone between Cys-546 and Gly-547) and the back-bone between Ile-24 and Glu-25 (included in the Dist-Grp calculation but not in Ene-Grp) contributed 5 kJ/mol. Assuming a dipole–dipole interaction (i.e., a r^{-3} dependence), the most long-ranged interaction is that with the amide group between Val-78 and the proton donor His-79 (4 kJ/mol at a distance of 4.4 Å); a similar interaction would still give a contribution of 3 kJ/mol at a distance of 5 Å or 1 kJ/mol at 8 Å distance.

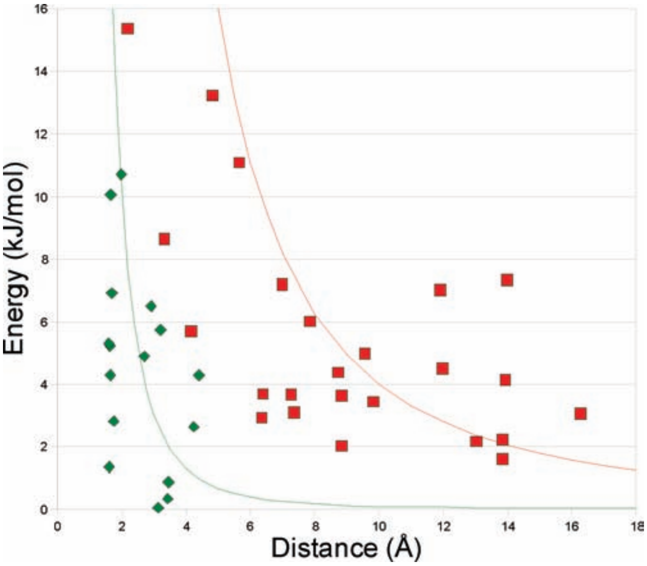

Figure 9. The relation between the absolute group contribution to the QM-only energy from each group (Ene-Grp calculation with COSMO) and its distance to the original QM system, divided into neutral (diamonds) and charged (squares) groups. The lines show the average r^{-3} and r^{-2} distance dependence of the interactions.

As expected, interactions with charged groups are both longer-ranged and stronger. The longest-ranged interaction is that with Asp-126 (7 kJ/mol at 14 Å distance). Assuming an r^{-2} distance dependence, such an interaction would still provide a 4 kJ/mol contribution at a distance of 20 Å. On the other hand, [Ni,Fe] hydrogenase is unusual in that it contains so many buried charged groups (~47, estimated from MD simulations); in most proteins, the great majority of charged groups are located on the surface. Both experimental and computational studies indicate that solvent-exposed surface charges have little influence on reaction energies.[11,24,25] Therefore, it is unlikely that residues at distances over 20 Å contribute significantly to reaction and activation energies in proteins.

However, our results give a quite pessimistic view of the QM-only approach for the study of enzyme reactions, even if a continuum solvation model is used. It is clear that it must entirely rely on cancellation of a large number of rather small contributions (up to 17 kJ/mol), which are quite long-ranged. Figure 4 shows that the energy may vary by up to 50 kJ/mol even if over 160 atoms are considered and the groups are added according to the QTCP energy contributions. Even if very large QM systems are used (~400 atoms), the energy will not be more accurate than 20 kJ/mol. Likewise, there is a 22 kJ/mol difference between the best Ene-Res and Ene-Grp results. Even worse, our results show that it is not enough to include in the QM system residues close to the active site; Figure 7 shows that this gives energies that are wrong by ~40 kJ/mol. Instead, neutral groups up to 4.5 Å and charged groups up to 16 Å from the active site contribute significantly to the energy difference. In traditional QM-only approaches, only groups directly connected or hydrogen-bonded to the active site are included. As is shown in Figures 3 and 9, this would exclude many important groups.

The great majority of previous QM-only studies of [Ni,Fe] hydrogenase have used only the metal ions and their first-sphere ligands as the QM system.[26] However, recent studies with extended QM systems have been published, including His-79,[27] His-79, Glu-25, Arg-476, and Asp-541,[26,28] the latter residues plus Asp-114,[29] and also Ser-499 and its backbone to Pro-498 (up to 137 atoms).[30] According to Table 1, these selections are wise, but it can be seen that there are many other groups both close to and farther away from the active site that have significant influence on the reaction energies.

Of course, it can be argued that buried charges in proteins typically come in ionic pairs so that the effect of the two groups may (partly) cancel. This is supported by Figure 3 (even if three partners of ionic pairs are missing, namely, Arg-S26, interacting with Glu-S22, Glu-294, interacting with Arg-85, and Asp-473 interacting with Arg-388). However, there are also more complicated chains, for example, the triplets of Arg-23, Asp-126, and His-538, as well as those of Asp-114, Arg-476, and Asp-541, making the cancellation less effective (the net effects are −7 and +4 kJ/mol for the two triplets). There are also some buried charges that do not have any charged groups in the neighborhood.

Conclusions

In this paper, we have studied how proton-transfer energies calculated with the QM-only approach converge with respect to the size of the QM system. We have tried two different schemes to decide what residues to include in the calculations. In the first, residues are simply added according to their distance to the QM system. In the second, we use approximate energy components of a QM/MM-FEP approach to guide the choice of groups. In a more practical use of the latter approach, a single QM/MM minimization of each state should be performed, and the contributions of each residue to the relative reaction energy can then be calculated at the MM level, using charges from a QM calculation for the active site.

Unfortunately, our results show that the results of the two approaches differ by 60 kJ/mol, even if ∼400 atoms are considered, which succinctly illustrates the problem of modeling an enzyme reaction by a QM-only approach. The difference is caused primarily by differences in the groups included in the two approaches, and the energy-based approach seems to give the more accurate results. However, it seems to be hard to devise an approach that provides accurate results without explicitly including a major part of the enzyme; a large number of groups have a significant influence on the energies (up to 17 kJ/mol), groups directly connected to the active site, neutral groups close (up to ∼5 Å from) to the active site, and charged groups both near and far (up to 16 Å away). This is in accordance with our previous QTCP results, which indicated that all residues up to 10 Å need to be included in the calculations before the calculated reaction energy converges to within 20 kJ/mol, even if solvent-exposed charges are excluded.[11]

Therefore, the best suggestion for a QM-only approach is probably to include only the minimum of residues needed for the chemical reaction in the QM calculations (in our case, this would be a 39 atom system, where the model of Glu-25 is removed from our standard QM system in Figure 1; it gives a QM energy of −8 kJ/mol in vacuum and 49 kJ/mol with COSMO). This will give the intrinsic reactivity of the active site. Inclusion of a continuum solvation model is recommended as a crude model of the surroundings and, at least in the present case, has a strong influence on the reaction energy, bringing it much closer to the true value than the vacuum energy. Inclusion of selected additional residues from the surroundings is not recommended because this will only bias the results without increasing the accuracy. For example, if only Arg-476 is added to the QM system, this will increase the reaction energy by +15 kJ/mol, but if the nearby Asp-541 and Asp-114 residues are also included, the effect is actually −6 kJ/mol in the opposite direction. The present results show that a wrong selection of residues can give results that are wrong by over 40 kJ/mol even if 400 atoms are included in the QM model. It must simply accepted that the QM-only approach is based on the hypothesis that the surrounding protein does not influence the reaction energy. Any unbiased investigation of the influence of the surroundings necessarily needs to be performed with an approach that takes a full account of all groups in the protein.

However, this does not mean that the QM/MM approach is without problem. On the contrary, it has severe convergence and sampling problems, as discussed in the Introduction. Moreover, polarization of the MM system is typically ignored, and junctions between the QM and MM systems may introduce serious artifacts.[5,6] Thus, improved methods to estimate the influence of a protein on the active site are needed.[31,32] The present calculations also point out the problems with buried charged groups in the protein. Our results indicate that they can have a large influence on the reaction energies, especially when they do not form ion pairs. To confirm that the results are relevant, it is necessary to check if these charged groups are stable inside of the protein and are not actually neutral at ambient pH. There are many methods to determine the pK_a of groups in proteins,[33−35] and such calculations should be routinely used in the setup of QM/MM calculations of proteins, at least when buried charges give important contributions to the energies. Among the charged groups discussed in this paper, the simple and fast PROPKA approach[35] predicts that Glu-S46, Glu-S75, Asp-63, and Arg-70 should be neutral. They have a net contribution of −13 kJ/mol to the proton-transfer energy.

In conclusion, the QM-only approach is useful to compare various tentative mechanisms of the active site. However, it is based on the hypothesis that the surrounding protein has no influence on the energies, which we show in this paper to be quite crude, at least for some types of reactions. Therefore, the estimated accuracy of the QM-only approach is appreciably lower than the normal estimates of 12−20 kJ/mol.[1,4] This problem will not be solved by simply making the QM system larger, e.g. by simply adding the closest residues, nor by adding a selected number of residues forming hydrogen bonds to the active site. On the contrary, this will only make the results biased. The only objective way to study the effect of the surroundings on the enzyme reaction is to explicitly model all residues at least up to 10−15 Å from the active site.

Acknowledgment. This investigation has been supported by grants from the Swedish research council and the Crafoord Foundation as well as by computer resources of Lunarc at Lund University.

Supporting Information Available: The raw data for the Dist-Res, Ene-Res, Dist-Grp, Ene-Grp, and Dist+Grp calculations are given as Tables S1−S5. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Siegbahn, P. E. M.; Blomberg, M. R. A. *Annu. Rev. Phys. Chem.* **1999**, *50*, 221–249.

(2) Siegbahn, P. E. M.; Blomberg, M. R. A. *Chem. Rev.* **2000**, *100*, 421–438.

(3) Siegbahn, P. E. M.; Borowski, T. *Acc. Chem. Res.* **2006**, *39*, 729–738.

(4) Himo, F. *Theor. Chim. Acta* **2006**, *116*, 232–240.

(5) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(6) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.

(7) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(8) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

(9) Gao, J. L.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(10) Rod, T. H.; Ryde, U. *J. Chem. Theory Comput.* **2005**, *1*, 1240–1251.

(11) Kaukonen, M.; Söderhjelm, P.; Heimdal, J.; Ryde, U. *J. Chem. Theory Comput.* **2008**, *4*, 985–1001.

(12) Volbeda, A.; Montet, Y.; Vernède, X.; Hatchikian, E. C.; Fontecilla-Camps, J. C *Int. J. Hydrogen Energy* **2002**, *27*, 1449–1461.

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

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

(15) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

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

(17) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–126.

(18) Klamt, A.; Schüürmann, J. *J. Chem. Soc., Perkin Trans. 2* **1993**, *5*, 799–805.

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

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

(21) Kaukonen, M.; Söderhjelm, P.; Heimdal, J.; Ryde, U. *J. Phys. Chem. B* **2008**, *112*, 12537–12548.

(22) Sevastik, R.; Himo, F. *Bioorg. Chem.* **2007**, *35*, 444–457.

(23) Hopmann, K. H.; Himo, F. *J. Chem. Theory Comput.* **2008**, *4*, 1129–1137.

(24) André, I.; Kesvatera, T.; Jönsson, B.; Åkerfeldt, K. S.; Linse, S. *Biophys. J.* **2004**, *87*, 1929–1938.

(25) Schutz, C. N.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 400–417.

(26) Siegbahn, P. E. M.; Tye, J. W.; Hall, M. B. *Chem. Rev.* **2007**, *107*, 4414–4435.

(27) Stadler, C.; De Lacey, A. L.; Monet, Y.; Volbeda, A.; Fontecilla-Camps, J. C. ; Conesa, J. C.; Fernández, V. M. *Inorg. Chem.* **2002**, *41*, 4424–4434.

(28) Siegbahn, P. E. M. *Adv. Inorg. Chem.* **2004**, *56*, 101–125.

(29) Siegbahn, P. E. M. *C. R. Chemie* **2007**, *10*, 766–774.

(30) Nilsson Lill, S. O.; Siegbahn, P. E. M. *Biochemistry* **2009**, *48*, 1056–1066.

(31) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.

(32) Söderhjelm, P.; Ryde, U. *J. Phys. Chem. A* **2009**, *113*, 617–627.

(33) Ullmann, G. M.; Knapp, E.-W *Eur. Biophys. J.* **1999**, *28*, 533–551.

(34) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523–544.

(35) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

JP9029024