Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 25 (2006) 481–486
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Fast empirical pK_a prediction by Ewald summation

Elmar Krieger [a,*], Jens E. Nielsen [b], Chris A.E.M. Spronk [a], Gert Vriend [a]

[a] Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen, Toernooiveld 1, 6525ED Nijmegen, The Netherlands
[b] School of Biomolecular and Biomedical Science, Centre for Synthesis and Chemical Biology, UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

Received 28 September 2005; received in revised form 21 February 2006; accepted 22 February 2006
Available online 27 April 2006

Abstract

pK_a calculations for macromolecules are normally performed by solving the Poisson–Boltzmann equation, accounting for the different dielectric constants of solvent and solute, as well as the ionic strength. Despite the large number of successful applications, there are some situations where the current algorithms are not suitable: (1) large scale, high-throughput analysis which requires calculations to be completed within a fraction of a second, e.g. when permanently monitoring pK_a shifts during a molecular dynamics simulation; (2) prediction of pK_as in periodic boundaries, e.g. when reconstructing entire protein crystal unit cells from PDB files, including the correct protonation patterns at experimental pH. Such in silico crystals are needed by 'self-parameterizing' molecular dynamics force fields like YASARA YAMBER, that optimize their parameters while energy-minimizing high-resolution protein crystals.

To address both problems, we define an empirical equation that expresses the pK_a as a function of electrostatic potential, hydrogen bonds and accessible surface area. The electrostatic potential is evaluated by Ewald summation, which captures periodic crystal environments and the uncertainty in atom positions using Gaussian charge densities. The empirical proportionality constants are derived from 217 experimentally determined pK_as, and despite its simplicity, this pK_a calculation method reaches a high overall jack-knifed accuracy, and is fast enough to be used during a molecular dynamics simulation. A reliable null-model to judge pK_a prediction accuracies is also presented.

Keywords: Protein pK_a prediction; Poisson–Boltzmann; Particle mesh Ewald; Crystal space; Force field optimization

1. Introduction

The prediction of pK_a values in proteins has made considerable progress over the last years [1,2]. The Poisson–Boltzmann equation (PBE) has become an important tool because it allows the calculation of the electrostatic potential in a heterogeneous solute–solvent system, taking into account dielectric boundaries and the ionic strength. Initial approaches to electrostatic calculations were based on rough approximations like spherical proteins [3]. The ability to solve the Poisson–Boltzmann equation for arbitrarily shaped proteins [4–6] cleared the path for a range of successful applications, such as studies of enzymatic activity [7], pH-dependent conformational changes [8] and protein stability [9–11]. These algorithms, however, are computationally expensive, and consequently led to the development of several simplified algorithms that avoid solving the PBE. Examples of these algorithms are the Debye–Hueckel approach [12] and the electrostatic screening functions [13,14].

pK_a calculations have always focused on proteins in their physiological environment, matching the experimental determination of pK_a values, which is also done in solution using NMR spectroscopy. However, the quality of pK_a calculations depends heavily on the availability of high resolution protein structures. NMR structures of sufficient resolution are often not available, and one is forced to predict solution pK_a values using X-ray structures. Much effort has been devoted to determining the regions of structural divergence, excluding residues involved in crystal contacts [15], optimizing X-ray structures [16] and incorporating information on protein flexibility [17].

1.1. The goal is pK_a prediction in protein crystals

The approach presented here has been developed due to a lack of solutions for a problem that appears paradoxical, given

Abbreviations: PME, particle mesh Ewald; PBE, Poisson–Boltzmann equation

* Corresponding author. Tel.: +31 24 3652248; fax: +31 24 3652977.
E-mail addresses: elmar@yasara.org, elmar@cmbi.ru.nl (E. Krieger).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.02.009

the facts mentioned above: the prediction of pK_a values in protein crystals. Because of the crystal packing interactions, these pK_as certainly differ from those measured in solution. The reason for addressing this problem becomes clear in view of recent developments in force field research. Thanks to the virtually unlimited resources provided by distributed computing systems like Models@Home [18], it became feasible to use complete proteins instead of small molecules as optimization targets when fitting the force field parameters [19]. This was done by randomly changing force field parameters and running simulations on a series of protein structures to see if the parameter changes would be beneficial. Obviously, the protein structures in the optimization set should be as realistic as possible, otherwise the force field might memorize features that are just structural artifacts. This can be achieved by taking high resolution X-ray structures and reconstructing the entire unit cell, including water molecules, counter ions and all solute hydrogens. The correct placement of polar hydrogens is especially important, and in addition to optimizing the hydrogen-bond network [20], this requires the pK_a values of all ionizable residues in the protein crystal and the pH at which the protein was crystallized. The force field parameters are then optimized in crystal space, so that all the interactions responsible for the experimentally observed structure can be considered, while converging at a force field like YAMBER [21]. Because crystal and solution environments obey the same laws of physics, the optimized force field can be used in both.

2. Ewald summation captures the periodic environment

Electrostatic calculations in periodic crystal systems are complicated by the infinite number of interactions. A clever way of making the problem tractable is Ewald summation [22], which allows the calculation of the potential due to the N particles in the unit cell and an infinite number of periodic replicas. The method combines a rapidly converging short-range term with a long-range component evaluated in reciprocal space [23]. If the reciprocal sum is calculated using a particle-mesh approximation, the resulting particle mesh Ewald (PME) algorithm [24] is considerably faster than the standard Ewald method. PME is part of almost every molecular dynamics program, and forms the basis for this work. However, we only use the reciprocal space portion, which provides the solution to Poisson's equation with periodic boundaries, Gaussian charge distributions and a single dielectric constant. By ignoring the short-range term and the associated damping of the reciprocal space term at short-range, we essentially remove the long-range attribute from the reciprocal space term: it now covers all distance ranges equally, and differs from Coulomb's law only by the use of Gaussian charge densities instead of localized point charges. Smeared-out Gaussians account for the uncertainty in atom positions (which also proved beneficial for the development of knowledge-based potentials [25]). Compared to the Poisson–Boltzmann equation, this approach however lacks two advantages: implicit counter ions and different dielectric constants for solvent and solute.

In an extensive optimization study, Demchuck and Wade [1] determined that the best dielectric constant for solvent exposed residues is close to the one of water (80), while the protein interior should be assigned a value in the range of 10–20. Since 20 differs from 80 only by a factor of 4, we hypothesized that a single global dielectric constant could suffice for accurate predictions, provided that some additional structural information was incorporated to account for the simplification.

3. The pK_a can be approximated as a function of electrostatic potential, hydrogen bonds and accessible surface

Using simplified physical considerations and some modeler's experience, we defined three rules of thumb for pK_a prediction. The first and partly the second rule have also been mentioned in a recent analysis of carboxyl pK_a values [26]:

- If an ionizable group is surrounded by negatively charged residues, corresponding to a negative electrostatic potential, protonation becomes easier, the pK_a increases. Similarly, if there are positively charged residues around, the pK_a decreases. As a first approximation, the pK_a shift is thus assumed to be proportional to the electrostatic potential.
- If an ionizable group accepts hydrogen bonds, the space to place a proton is reduced, protonation becomes harder, and the pK_a decreases. If after protonation, the group can donate a bond, protonation is favorable, the pK_a increases.
- If a group accepts hydrogen bonds and is buried, the pK_a is decreased even further, because the side-chain cannot facilitate protonation by moving to a different conformation where it does not receive hydrogen bonds. If a buried group can donate a hydrogen bond after protonation, the pK_a increases, because there is no space for water molecules that could ease the energetic cost of two hydrogen-bond acceptors facing each other.

These three assumptions were fused into an empirical equation relating the pK_a of a residue with the electrostatic potential, the number of hydrogen bonds and the accessible surface area:

$$\mathrm{p}K_a = \mathrm{Model}\,\mathrm{p}K_a + \sum_{\text{Ionizable atoms}} [-A \times \mathrm{Ewald}E_i + B \times \mathrm{HB}_i] + \mathrm{Sign}(\mathrm{HBSum}) \times C \times \mathrm{SurfaceLoss} \quad (1)$$

In this equation, *Model* pK_a is the standard pK_a value of a certain residue type, EwaldE_i is the reciprocal space portion of the Ewald energy of a charge +1 at the location of the ith ionizable atom in the residue (in kcal/mol), HB$_i$ is the difference between (potentially) donated and accepted hydrogen bonds at the ith atom, HBSum is the sum over all HB$_i$, and SurfaceLoss is the loss of accessible surface area of the side-chain with respect to a fully exposed state. A, B and C are empirical proportionality constants. The four unknown parameters *Model* pK_a, A, B and C are globally optimized for each amino acid type so that the RMSD between predicted and