pepKalc

Tamiola, Kamil; Scheek, Ruud M.; van der Meulen, Pieter; Mulder, Frans A. A.

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/bty033

Publication date:
2018

Bioinformatics, 34(12), 2018, 2053–2060
doi: 10.1093/bioinformatics/bty033
Advance Access Publication Date: 22 January 2018
Original Paper

Structural bioinformatics

pepKalc: scalable and comprehensive calculation of electrostatic interactions in random coil polypeptides

Kamil Tamiola[1,2,*], Ruud M. Scheek[2], Pieter van der Meulen[2] and Frans A. A. Mulder[2,3,*]

[1]Peptone – The Protein Intelligence Company, Amsterdam 1001AM, The Netherlands, [2]Department of Molecular Dynamics, GBB, University of Groningen, Groningen 9747AG, The Netherlands and [3]Department of Chemistry and Interdisciplinary Nanoscience Center iNANO, Aarhus University, Aarhus 8000, Denmark

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 22, 2017; revised on December 17, 2017; editorial decision on January 8, 2018; accepted on January 19, 2018

Abstract

Motivation: Polypeptide sequence length is the single dominant factor hampering the effectiveness of currently available software tools for *de novo* calculation of amino acid-specific protonation constants in disordered polypeptides.

Results: We have developed **pepKalc,** a robust simulation software for the comprehensive evaluation of protein electrostatics in unfolded states. Our software completely removes the limitations of the previously reported Monte-Carlo approaches in the computation of protein electrostatics by using a hybrid approach that effectively combines exact and mean-field calculations to rapidly obtain accurate results. Paired with a modern architecture GPU, **pepKalc** is capable of evaluating protonation behavior for an arbitrary-size polypeptide in a sub-second time regime.

Availability and implementation: http://protein-nmr.org and https://github.com/PeptoneInc/pepkalc

Contact: kamil@peptone.io or fmulder@chem.au.dk

1 Introduction

Protonation is a ubiquitous and important process in biology. Protein folding, ligand recognition, enzyme catalysis, membrane potentials and the energetics of cells depend on ionization and proton transfer (Kumar and Nussinov, 2002). Thus great theoretical and experimental effort have been devoted to the elucidation of the intricate workings of protonation-related electrostatics at the molecular level (Alexov *et al.*, 2011; Hass and Mulder, 2015; Wallerstein *et al.*, 2015).

In their seminal work, Tanford and Kirkwood introduced a theoretical framework for the treatment of protein electrostatics in solution (Tanford and Kirkwood, 1957). The 'hard-sphere' model of protein electrostatics, proposed by Tanford and Kirkwood, was further improved by an inclusion of a mean-field approximation of pair-wise interactions in a polypeptide chain (Tanford, 1961). Furthermore, the developments in X-ray crystallography and advancements in computational structural biology opened a path to a new class of protein electrostatics models. Bashford and Karplus showed that a continuum electrostatics with distance-distributions derived from the structural ensembles of proteins could be used to compute the site-specific protonation behavior in folded polypeptides (Bashford and Karplus, 1990). The continuum model got further improved by an inclusion of protein flexibility and local dynamics gauged from molecular dynamics sampling (You and Bashford, 1995). Although significant progress in the theoretical treatment of the protonation behavior of folded polypeptides has been made, a complete characterization of electrostatic interactions in folded proteins still remains a formidable task. Incomplete treatment of protein dielectric properties and rudimentary conformational sampling were implicated as crucial factors hampering the elucidation of protonation behavior at amino acid level. The experimental study of RNase T1 by Bombarda and Ullmann elegantly demonstrates the global

challenges in predicting and explaining the intricate protonation behavior of amino acid residues in folded proteins (Bombarda and Ullmann, 2010). Their study suggested that the interactions between protonatable residues in proteins can help to maintain the energy required to protonate a site in the protein nearly constant over a wide pH range, suggesting an existence of coupled networks of reside-residue interactions.

In addition, numerous successful protonation-state prediction studies have been reported for purposefully unfolded and intrinsically disordered proteins (IDPs) (Elcock, 1999; Geist *et al.*, 2013; Zhou, 2001, 2002a,b). It was demonstrated that elucidation of electrostatic interactions in polypeptides devoid of canonical structural features benefits greatly from the adoption of a Gaussian-chain, as a model for residue-to-residue distances in a structurally disordered state (Zhou, 2001). Unfortunately, the Monte-Carlo based free energy integration, presented in Zhou (2002a) was successful only at reproducing experimental data for small-size polypeptides. The predictions of the protonation behavior for proteins with more than twenty residues were unattainable given the need for representative sampling of all possible configurations (Zhou, 2001). Motivated by the simplicity and the effectiveness of the Gaussian-chain model, we decided to extend its applicability to an arbitrary peptide of any sequence length and complexity.

We present here **pepKalc,** a robust software for the calculation of protonation constants in an unfolded polypeptide of arbitrary length. Combining the simplicity of the Gaussian-chain model for the disordered state with the robustness and speed of a hybrid mean-field approximation treatment (Gilson, 1993; Yang *et al.*, 1993), all the relevant pair-wise electrostatic interactions can be calculated within seconds. Finally, we demonstrate how **pepKalc** can help rationalize experimental pKa shifts in an intrinsically disordered protein.

2 Theory

2.1 The general theory

The association reaction to form an unfolded polypeptide chain in a specific protonation state x at position i can be schematically represented as,

$$(x_i)H^+ + X_i^{q_0} \iff HX_i^{q_0+x_i} \tag{1}$$

where $X_i^{q_0}$ is the fully deprotonated form of a an amino acid residue at the position i, with the unit charge of q_0. H represents the protons, and $HX_i^{q_0+x_i}$ is the site i of polypeptide in protonation state x. Correspondingly, the collective protonation state of an arbitrary unfolded polypeptide with N titratable sites can be expressed in terms of a vector x, whose elements, x_i, take on values of 0 or 1 according to whether the corresponding site i is unprotonated or protonated, respectively.

$$x = \{x_1, x_2, x_3 \ldots x_N\} \quad x_i = \begin{cases} 0 & \text{if deprotonated} \\ 1 & \text{if protonated} \end{cases} \tag{2}$$

Residues within the polypeptide chain can be classified based on their residual charge in the deprotonated form into two distinct groups, acidic $A \in \langle C-terminus, C, D, E, Y\rangle$ with $q_0 = -1$, and basic $B \in \langle N-terminus, H, K, R\rangle$ and $q_0 = 0$ with the intrinsic and amino-acid specific pK_a values listed in Table 1. The explicit treatment of protonation behavior of an arbitrary unfolded protein with multiple electrostatically interacting sites, can be demonstrated on an example of a fully deprotonated polypeptide, composed of basic residues BBB. In the presence of a single proton, the protonation reaction of BBB at its first available site,

$$BBB + H^+ \overset{K_1}{\iff} B^+BB \tag{3}$$

yields a partially protonated polypeptide B^+BB, whose change in free energy of protonation relies solely on the intrinsic dissociation constant K_1.

$$\Delta G = -RT\,ln(K_1) \tag{4}$$

Under such conditions, the concentration of protonated polypeptide B^+BB can be expressed as,

$$[B^+BB] = [BBB]\frac{[H^+]}{K_1} = [BBB]\,10^{pK_1-pH} \tag{5}$$

where pK_1 is the intrinsic pK_a value of the first residue in the polypeptide. An inclusion of a second proton,

$$B^+BB + H^+ \overset{K_2}{\iff} B^+B^+B \tag{6}$$

causes a significant change in the free energy of the protonation reaction due to charge–charge interactions between the protonated sites 1 and 2,

$$\Delta G = \Delta G_1 + \Delta G_2 + G_{12} \tag{7}$$

where ΔG_1 and ΔG_2 are the self-association free energy changes, and G_{12} denotes an electrostatic contribution to the overall free energy change due to charge–charge interactions. In the presented approach, the electrostatic charge–charge interactions are assumed to follow the linearized Debye-Hückel equation (Tanford, 1961), adopted for a solvated polymer chain in a solution of known dielectric constant ϵ, and ionic strength I. Following the approach of Zhou, the distance between the charged residues i and j, that is $r_{i,j}$, is assumed to be sampled from the distribution for a Gaussian Chain $p(r)$,

$$p(r_{i,j}) = 4\pi r_{i,j}^2\left(\frac{3}{2\pi d_{i,j}^2}\right)^{\frac{3}{2}}\exp\left(-\frac{3r_{i,j}^2}{2d_{i,j}^2}\right) \tag{8}$$

where $d_{i,j}$ is the root-mean-square (RMSD) distance between the interacting sites, approximated by the empirical relation,

$$d_{i,j} = b\sqrt{l_{i,j}} + s \tag{9}$$

in which, $l_{i,j}$ is the number of peptide bonds separating the interacting i, j residues, b is the effective residue separation, and s represents the shift distance due to side chain. These model values are adjustable parameters, and are taken to be $b = 7.5$ Å and $s = 5$ Å (Zhou, 2002b). An assumption is made that $p(r)$ is independent of

Table 1. Model compound pK_0 values (Hass and Mulder, 2015)

Residue type	Model pK_*
N-Terminus	8.0
C-Terminus	3.6
D	4.0
E	4.4
H	6.3
C	8.3
Y	9.6
K	10.4
R	13.5

temperature, ionic strength and pH. The mean electrostatic interaction energy between sites i, j, that is G_{ij}, can be computed from,

$$G_{ij} = 332\sqrt{6/\pi}\,\frac{1-\sqrt{\pi}\,\beta\, e^{\beta^2}\ \mathrm{erfc}(\beta)}{\epsilon\, d_{i,j}} \tag{10}$$

where $\beta = \frac{\kappa d_{i,j}}{\sqrt{6}}$, $\kappa = \sqrt{\frac{8\pi I e^2}{\epsilon k_B T}}$, I is ionic strength, T is temperature, ϵ is the dielectric permitivity of a solvent and erfc(β) is the complementary error function (Zhou, 2002b). Consequently, all possible configurations of the model polypeptide BBB arising from the complete protonation, can be expressed in terms of $[BBB]$ and the concentrations of the free proton ligand, $[H^+]$, as follows,

$$\begin{aligned}
[BBB^+] &= [BBB]\frac{[H^+]}{K_3}\\
[BB^+B] &= [BBB]\frac{[H^+]}{K_2}\\
[BB^+B^+] &= [BBB]\frac{[H^+]^2}{K_2K_3p_{23}}\\
[B^+BB] &= [BBB]\frac{[H^+]}{K_1}\\
[B^+BB^+] &= [BBB]\frac{[H^+]^2}{K_1K_3p_{13}}\\
[B^+B^+B] &= [BBB]\frac{[H^+]^2}{K_1K_2p_{12}}\\
[B^+B^+B^+] &= [BBB]\frac{[H^+]^3}{K_1K_2K_3p_{12}p_{13}p_{23}}
\end{aligned} \tag{11}$$

where p_{12}, p_{13}, p_{23}, represent an increase in the free energy of a particular configuration, due to the electrostatic interactions between the like (positive) charges at sites 1–2, 1–3 and 2–3, respectively. The pair-wise electrostatic interaction parameters p_{ij} are dimensionless and computed from

$$\log(p_{ij}) = \frac{4184.0}{RTln(10)}G_{ij} \tag{12}$$

where G_{ij} represents electrostatic repulsion introduced in Equation 10. Consequently, the electrostatic interaction Hamiltonian of a simulated system can be represented as W_{ij}.

$$W_{ij} = \begin{bmatrix} 0 & \log(p_{12}) & \cdots & \log(p_{1N}) \\ \log(p_{21}) & 0 & \cdots & \log(p_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ \log(p_{N1}) & \log(p_{N2}) & \cdots & 0 \end{bmatrix} \tag{13}$$

Furthermore, the factor $\frac{[H^+]^2}{K_iK_jp_{ij}}$, can be expanded to $10^{pK_i+pK_j-2pH-\log(p_{ij})}$ in our symbolical framework.

For reasons that will become clear below we shall express the concentrations of all configurations relative to that of the fully deprotonated configuration. However, each deprotonated acidic residue in a polypeptide introduces a negative charge in this reference state, making this case different from the all-basic case above. We shall illustrate our proposed approach with the tetrapeptide $[BBBA^-]$. Two pathways connect the doubly protonated configuration $[BBB^+A]$ with the fully deprotonated state $[BBBA^-]$,

$$BBB^+A \overset{K_3}{\Longleftrightarrow} BBBA + H^+ \overset{K_4}{\Longleftrightarrow} BBBA^- + H^+ + H^+ \tag{14}$$

and,

$$BBB^+A \overset{K_4p_{34}}{\Longleftrightarrow} BBB^+A^- + H^+ \overset{\frac{K_3}{p_{34}}}{\Longleftrightarrow} BBBA^- + H^+ + H^+ \tag{15}$$

resulting in the overall equation,

$$[BBB^+A] = [BBBA^-]\frac{[H^+]}{K_3}\frac{[H^+]}{K_4} \tag{16}$$

The second pathway, Equation 15, suggests an alternative point of view, similar to the all-basic case above,

$$[BBB^+A] = [BBBA^-]\frac{[H^+]}{\frac{K_3}{p_{34}}}\frac{[H^+]}{K_4p_{34}} \tag{17}$$

Substitution of $\frac{K_3}{p_{34}}$ by $K_3' = \frac{K_3}{p_{34}}$ yields,

$$[BBB^+A] = [BBBA^-]\frac{[H^+]}{K_3'}\frac{[H^+]}{K_4}\frac{1}{p_{34}} \tag{18}$$

This view implies a correction of the intrinsic pK_a of residue 3 $(K_3 \rightarrow K_3')$ for the negative charge at position 4. Thus we achieve our goal, namely that in the case of double protonation of a basic and an acidic residue one repulsive interaction term $\left(\frac{1}{p_{34}}\right)$ is introduced, like in the case of protonation of two basic residues, expressed in Equation 11. Applying this reasoning to the other configurations allows us to represent all 16 configurations of this tetrapeptide relative to the concentration of the fully deprotonated state $[BBBA^-]$ in a homogeneous manner,

$$\begin{aligned}
[BBB^+A^-] &= [BBBA^-]\frac{[H^+]}{K_3'}\\
[BB^+BA^-] &= [BBBA^-]\frac{[H^+]}{K_2'}\\
[BB^+B^+A^-] &= [BBBA^-]\frac{[H^+]^2}{K_2'K_3'p_{23}}\\
[B^+BBA^-] &= [BBBA^-]\frac{[H^+]}{K_1'}\\
[B^+BB^+A^-] &= [BBBA^-]\frac{[H^+]^2}{K_1'K_3'p_{13}}\\
[B^+B^+BA^-] &= [BBBA^-]\frac{[H^+]^2}{K_1'K_2'p_{12}}\\
[B^+B^+B^+A^-] &= [BBBA^-]\frac{[H^+]^3}{K_1'K_2'K_3'p_{12}p_{13}p_{23}}\\
[BBBA] &= [BBBA^-]\frac{[H^+]}{K_4}\\
[BBB^+A] &= [BBBA^-]\frac{[H^+]^2}{K_3'K_4p_{34}}\\
[BB^+BA] &= [BBBA^-]\frac{[H^+]^2}{K_2'K_4p_{24}}\\
[BB^+B^+A] &= [BBBA^-]\frac{[H^+]^3}{K_2'K_3'K_4p_{24}p_{34}p_{23}}\\
[B^+BBA] &= [BBBA^-]\frac{[H^+]^2}{K_1'K_4p_{14}}\\
[B^+B^+BA] &= [BBBA^-]\frac{[H^+]^3}{K_1'K_2'K_4p_{14}p_{24}p_{12}}\\
[B^+BB^+A] &= [BBBA^-]\frac{[H^+]^3}{K_1'K_3'K_4p_{14}p_{34}p_{13}}\\
[B^+B^+B^+A] &= [BBBA^-]\frac{[H^+]^4}{K_1'K_2'K_3'K_4p_{14}p_{24}p_{34}p_{12}p_{13}p_{23}}
\end{aligned} \tag{19}$$

In the light of the explicit formalism presented in Equation 19, the

individual fractional protonation states $\langle x_i \rangle$ at each pH, can be calculated from

$$\langle x_i \rangle (pH) = \frac{\sum [XH_i]}{\sum [XH]} \tag{20}$$

where $\sum [XH]$ denotes the summation over all 2^N possible protonation states a model polypeptide with N electrostatically interacting sites would have and the index i limits the sum to states with site i protonated. Ultimately, the pK_a values of titratable residues are estimated by fitting the pH dependence of $\langle x_i \rangle$ to the Hill equation,

$$\log \frac{\langle x_i \rangle}{1 - \langle x_i \rangle} = n_i (pK_a - \mathrm{pH}) \tag{21}$$

where n_i is the Hill parameter, under the assumption that all simulated sites adapt their equilibrium protonation states during the titration event (Bombarda and Ullmann, 2010).

2.2 The hybrid mean field method

An explicit numerical integration over all possible protonation-state combinations is required to solve Equation 20, yielding a typical $O(2^n)$ scaling problem. To mitigate the scalability issue, a hybrid mean field approach (Gilson, 1993; Yang *et al.*, 1993) was developed. Algorithm 1, provides an overview of our approach. In the proposed methodology, given a polypeptide with N sites an explicit treatment of all plausible protonation states is applied only to a window of interacting residues w, defined as,

$$w = 2\phi + 1 \tag{22}$$

where ϕ is the interaction cutoff parameter. The calculation window w is recursively iterated α_{max} times over $i \in \langle \phi + 1, N - \phi \rangle$. In each cycle α corrected pK_a(s) are calculated for residue $i \in \langle 1, N \rangle$, following the explicit procedure given by Equations 20 and 21, but now implying only residues within the range $i - \phi, i + \phi$. In addition to the explicit corrections for negative (unit) charges on acidic residues within that window (see above), also pK_a-corrections for charges q_j outside the window are taken into account, according to,

$$W_i = \sum_j q_j \, W_{ij} \tag{23}$$

assuming that averaged (fractional) charges can be used for these. These averaged charges q_j follow directly from the fractional protonations $\langle x_j \rangle$ calculated in the previous cycle (In the first cycle they are calculated from tabulated pKa values (Table 1) assuming Hill parameters of $n = 1.0$). This is repeated for all pH values in the chosen range before moving to the next residue (and the next window w). A new cycle (α) is started if the new pK_a(s) differ too much from those calculated in the previous cycle.

Algorithm 1 Hybrid Mean Field Approximation

Require:
protein sequence with N interacting sites,
residue cutoff ϕ,
number of supercycles α: α_{max}
Ensure:
acidity constants pK_a,
Hill paramters n,
$W_{ij} \leftarrow \log(p_{ij}) \leftarrow G_{ij}$ ▷ position-dependent interaction energies
for α **do**
  for $i \in \langle 1, N \rangle$ **do**
    for pH **do**
      if $\alpha = 1$ **then** ▷ the first cycle
        $q_j \leftarrow (pK_a, n_i = 1.0)$ ▷ use intrinsic pK_a(s)
      else
        $q_j \leftarrow \langle x_j \rangle$ ▷ use values from previous iteration
      end if
      if $j < i - \phi \wedge j > i + \phi$ **then** ▷ residues outside w
        $W_i \leftarrow \sum_j q_j W_{ij}$ ▷ according to Equation 23
      end if
      $X_i, X \leftarrow W_{ij}, W_i$
      $\langle x_i \rangle \leftarrow \frac{X_i}{X}$ ▷ Equation 20
    end for
    pK_a, n for residue $i \leftarrow \langle x_i \rangle$ ▷ Equation 21
  end for
  if $\alpha = \alpha_{max}$ **then**
    Stop
  end if
end for

3 Materials and methods

3.1 Software implementation

pKa-calc software is available as a stand-alone, interactive HTML5 application at http://protein-nmr.org and command-line utility at https://github.com/PeptoneInc/pepkalc. The tool was developed using Python 2.7 (van Rossum, 1995). The self-consistent numerical integration routines, were implemented using high performance nVidia CUDA libraries, according to single-data-multiple-instruction (SIMD) paradigm. **pKa-calc** supports deployment in virtual machine environments (cloud servers) with concurrent thread technology and transparent execution on nVidia graphical processing units (GPU). The command-line utility version of **pepKalc,** can be readily containerized using Docker engine (Hykes, 2018) and deployed as a ultra-scalable application with an integrated support for remote-procedure-call (RPC) servers.

3.2 Performance testing and deployment

All performance benchmarks were performed on Apple Mac Pro computer (Mid 2012 model), equipped with two 2.4 GHz 6-Core Intel Xenon E5645 processors, 16GB 1333 MHz DDR3 ECC memory, nVidia GeForce GTX680 4 GB GPU, operating under the control of OSX 10.11.4 64-bit system with nVidia CUDA v7.5 architecture drivers. Multi-GPU scalability benchmarks were performed on nVidia DGX-1 supercomputing node.

4 Results and discussion

The performance of **pepKalc** was assessed. We have benchmarked the numerical convergence of the tool as a function of input parameters and the time required for the completion of benchmark simulations. Subsequently, we have evaluated the predictive power of **pepKalc** for the strong electrostatic-coupling case. Ultimately, we have compared experimental protonation constants for intrinsically disordered human alpha-synuclein against **pepKalc** predictions.

4.1 Numerical convergence

One of the most important issues with the self-consistent numerical integration routines is an estimation of the minimal number of iterations required for the method to reach a satisfactory level of numerical convergence (Gear *et al.*, 1967). A protocol was devised to

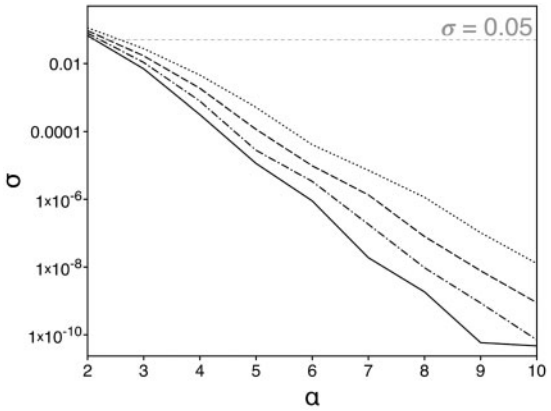

Fig. 1. The assessment of numerical convergence of **pepKalc**. The standard deviation of the residuals of the pK_a values for a D_{40} polypeptide with $N = 40$. The standard deviation σ is plotted on a logarithmic scale against the number of the integration super-cycles α, for four window sizes: $w = 3$ (dotted line $\cdots$), $w = 5$ (dashed line $--$), $w = 7$ (semi-dashed line $-$) and $w = 9$ (continuous line). The threshold line $\sigma = 0.05$ is marked on the plot with gray. The remaining simulation parameters were: $T = 298.15K$, $I = 0.1M$ and $\epsilon = 78.5$

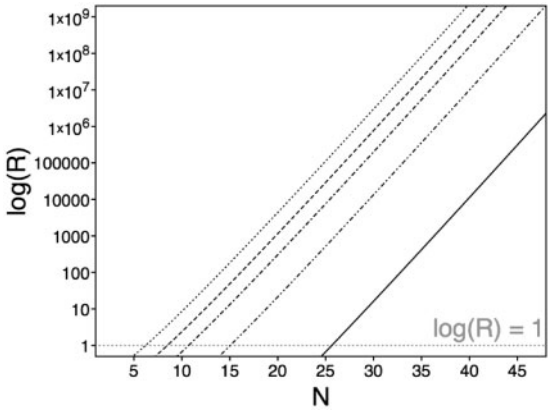

Fig. 2. The theoretical performance gain of hybrid mean field approach over an explicit calculation, computed from Equation 27. The performance advantage is expressed on a logarithmic scale $\log(R)$, as a function of the number of interacting sites N. The hybrid mean field method calculation was performed with $\alpha = 3$ super-cycles. The performance curves are computed for four cutoff parameter ϕ values: $\phi = 1$ (dotted line $\cdots$), $\phi = 2$ (dashed line $--$), $\phi = 3$ (semi-dashed line $-\cdot$) and $\phi = 4$ (continuous line). The dashed gray line corresponds to a performance threshold ($\log(R) = 1$), above which computational gain is expected when using **pepKalc** tool

determine the minimum, acceptable number of iterations for **pepKalc** tool that lead to numerical convergence. The test consisted of a computation of pK_a values for a $N = 40$ residue polypeptide, composed solely of aspartate residues (D_{40}). The residue-specific pK_a values, predicted every iteration j, were stored in a form of a row-ordered matrix K. The numerical convergence criterion was defined as a standard deviation σ of a residual difference between the computed pK_a values in the consecutive iterations j and $j - 1$. The following steps were taken in the estimation of the numerical convergence of **pepKalc** tool:

(1) a computation of the residuals vector $\Delta p\vec{K}_{a|j}$, for every iteration j,

$$\Delta p\vec{K}_{a|j} = \vec{K}_j - \vec{K_{j-1}} \iff j \in \langle 2, \alpha \rangle \tag{24}$$

where $\vec{K}_j$ is the row-vector of predicted pK_a values after iteration j.

(2) an estimation of the standard deviation $\sigma(j)$, for every iteration j according to,

$$\sigma(j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta p\vec{K}_{a|j}(i) - \mu\right)^2} \tag{25}$$

where,

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\Delta p\vec{K}_{a|j}(i) \wedge j \in \langle 2, \alpha \rangle \tag{26}$$

Figure 1 summarizes the outcome of the convergence estimation procedure. Four, ten super-cycle ($\alpha = 10$) simulations were performed with variable interaction window sizes w. It was found the calculations reached numerical convergence below a threshold of $\sigma = 0.05$ already after $\alpha = 3$ cycles, irrespectively of the computational window w size. An increase in the number of computational iterations ($\alpha \geq 6$) yielded even lower standard deviation levels of computed residuals, surpassing the value of 0.0001 for all computed scenarios.

4.2 Performance gain factor

The relative performance of the hybrid mean field method in **pepKalc** software with respect to an exact result from an exhaustive calculation variant was assessed. A performance gain factor R was introduced. The gain is defined here as a ratio of the number of steps required for a numerically convergent simulation in an explicit approach, to the overall number of iterations in the hybrid mean field method algorithm,

$$R = \frac{2^{N-2\phi-1}}{\alpha N} \tag{27}$$

where N is the number of interacting sites, ϕ is the size of the hybrid mean field method cutoff, and α is the number of self-consistency integration super-cycles. For the sake of an example, let us consider a protonation state prediction for a 40-site polypeptide. An $N = 40$ calculation in an explicit approach would require approximately 4TB of RAM memory in order to accommodate the protonation state matrix, with 2^{40} state combinations, stored as 32-bit precision floats. Subsequently, $2^{40} = 1099511627776$ integration cycles would be required to compute numerically convergent solution to Equation 20. The very same computational problem, given $\alpha = 3$ integration super-cycles, with a cutoff parameter of $\phi = 2$, requires approximately, 480KB of RAM memory, to store 2^5 protonation-sate combinations, and $3 \cdot 40 \cdot 2^{2\cdot 2+1} = 7680$ integration cycles, when computed with **pepKalc** software. The expected performance gain R due to switching from an explicit computation to reduced site-approximation is $\frac{2^{35}}{(3\cdot 40)} \approx 285000000$. Figure 2 demonstrates the relative gain in the computational performance, expressed as a $\log(R)$, against the number of interacting sites N.

A meaningful, that is $\log(R) \gg 1$, performance enhancements can only be achieved if the number of the interacting sites N significantly exceeds the size of the computational cutoff (ϕ). In the case of $\phi = 1$ computational performance improvements are expected for proteins with $N > 7$, reaching 100-fold speedup for a polypeptide with fourteen interacting sites ($N = 14$). A clear performance shift can be observed in the theoretical performance curves for $\phi = 2, 3, 5$ and 10, advising the use of small computational window sizes in most cases of practical interest.

4.3 Calculation time benchmarks

A series of **pepKalc** execution-time benchmarks was performed. The multi-processor scalability and the numerical performance of the

GPU implementation were assessed. Table 2 summarizes the outcome of the testing procedure for $D_{40}, N = 40$ polypeptide with three $\alpha = 3$ integration super-cycles, but variable calculation cutoff ϕ. Each benchmark simulation was executed fifty times and concluded with the calculation of the mean simulation time. As evidenced by the benchmark data, **pepKalc** retains near-linear scalability across the tested range of CPU cores. However, a GPU-accelerated implementation of **pepKalc** offers a remarkable two-orders of magnitude increase in the computational efficiency with respect to the CPU code, achieving 60x speedup for the most complex $\phi = 6$ computational scenario.

4.4 Strong electrostatic coupling

The predictive power of **pepKalc** in the strong-electrostatic coupling scenario was assessed. The theoretical protonation curves for *DDD* tripeptide were simulated. Five predictions were made with variable dielectric constant ϵ values. Figure 3 depicts the outcome of the calculations.

Low dielectric constant simulation $\epsilon = 2$ reveals very interesting protonation behavior, elegantly demonstrating the power of **pepKalc** in predicting a convoluted electrostatic coupling scenario. In the case of very low dielectric screening, a multi-site, correlated deprotonation of terminal D_1, and D_3 residues takes place whilst D_2 happens to retain its neutral character and reverts to its nearly neutral form at pH 10. Doubling of a dielectric constant, that is $\epsilon = 4$, yields complete deprotonation of D_2 at pH 12, smoothing out the multi-site character of curves for terminal residues D_1 and D_3. A typical, sigmoidal deprotonation behavior can be observed for simulations with $\epsilon > 20$, whereas the case of $\epsilon = 80$ resembles deprotonation events expected from Equation 21, although the Hill parameter does not have a unique interpretation for this particular case of weak-coupling (Bombarda and Ullmann, 2010; Lindman *et al.*, 2006).

Table 2. Simulation execution time benchmarks in seconds [s], reported for D_{40} polypeptide $N = 40$ with variable cutoff ϕ size

ϕ	1 CPU	2 CPUs	4 CPUs	GPU*
2	3.4	1.8	0.9	0.04
3	4.6	2.6	1.3	0.06
4	8.4	4.4	2.3	0.12
5	28.3	14.9	7.9	0.44
6	125	69.4	32.9	2.05

*Using nVidia GeForce 680 GTX card with 4GB DDR and 1536 computational threads.

4.5 Site-specific protonation behavior of intrinsically disordered alpha-synuclein

Human alpha-synuclein is a 14.5 kDa intrinsically disordered protein expressed predominantly at the presynaptic terminals of brain neurons, (George, 2002). It is known, that the misfolding of alpha-synuclein leads to the formation of fibrillar cytoplasmic aggregates, (Wood *et al.*, 1999) often referred to as Lewy bodies, which are defining characteristic of Parkinson's disease, (Cookson, 2005). Because of its pivotal role in the etiology of neurodegeneration, great effort has been devoted to its biophysical characterization. Human alpha-synuclein contains 140 amino acids, of which 46 have the capacity to act as acids or bases: 15 Lys, 1 His, 4 Tyr, 6 Asp, 18 Glu and the N- and C-termini. Croke *et al.* reported an experimental site-specific characterization of electrostatic interactions in alpha-synuclein using solution-state NMR spectroscopy. The pK_a values for all 26 sites that ionize below pH 7 were characterized using 2 D 1H–15N HSQC and 3 D C(CO)NH NMR experiments.

We have used **pepKalc** to predict the pK_a constants of ionizable residues in human alpha-synuclein adopting the experimental conditions found in (Croke *et al.*, 2011) as model input parameters. The calculation was performed for 46 independently interacting sites over the full pH range; thereby accurately treating pH-dependent protonation in the full-length polypeptide. It should be noted, the prediction of the electrostatic interactions with the similar level of numerical complexity is virtually impossible using the Monte-Carlo approach published in (Zhou, 2002b). We assumed in our calculations the experimental salt-free samples could be approximated by low ionic strength ($I = 0.001M$). Figure 4 depicts a comparative analysis of experimentally and computationally derived differences between the observed pK_a values and the reference constants pK_0 listed in Table 1.

As demonstrated in Figure 4 the pK_a values predicted by **pepKalc** follow the general trend found in the experimental study by (Croke *et al.*, 2011). The root mean square deviation (RMSD) of the computed pK_a constants with respect to the experimental values was used to asses the predictions. The simulations for the low ionic strength solution $I = 0.001M$ yielded an RMSD of 0.65, whereas calculations for $I = 0.150M$, resulted in much better agreement with the experimental values reflected by RMSD of 0.15.

The observed discrepancy in the estimation of ΔpK_a for human alpha-synuclein is expected, based on a massive body of

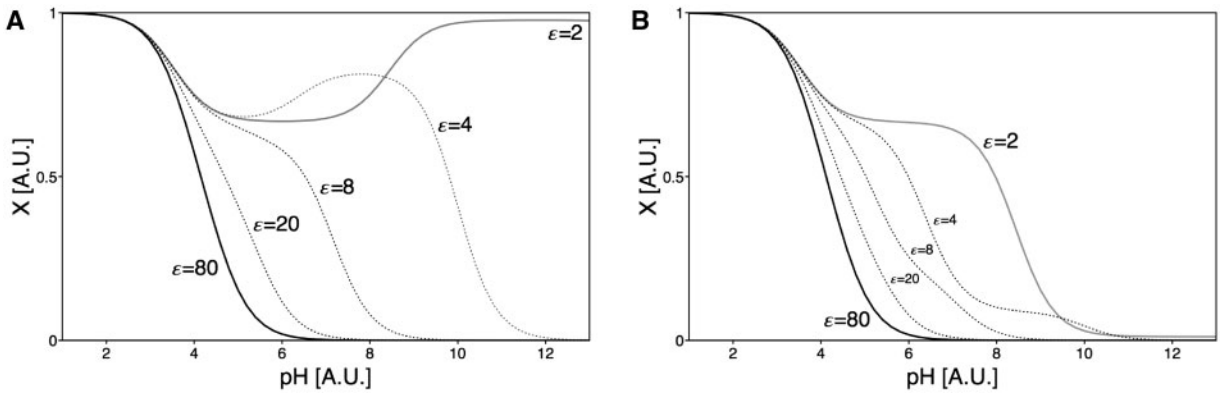

Fig. 3. (**A**) Protonation curves for D_1 and D_3, (**B**) D_2 residues of *DDD* polypeptide simulated with variable dielectric constant ϵ. The remaining simulation parameters are, $\alpha = 3$, $T = 298.15K$ and $I = 0.1M$

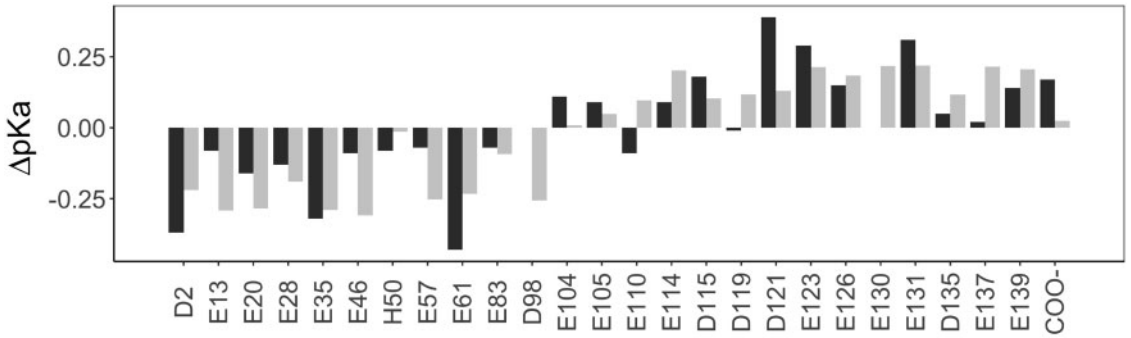

Fig. 4. Comparative analysis of experimental (black) and predicted (grey) ΔpK_a values for intrinsically disordered human α-synuclein. Calculation parameters were derived from Croke et al., $T = 283.15K$, $\epsilon = 83.83$ and $I = 0.150M$

experimental work done for this particular polypeptide. Growing experimental evidence from small angle X-ray scattering (SAXS) (Receveur-Brechot and Durand, 2012), solution state NMR spectroscopy (Nielsen and Mulder, 2016; Pierattelli and Felli, 2015) and empirical modeling (Jha *et al.*, 2005) studies, suggests intrinsically disordered proteins cannot be represented as homopolymers devoid of canonical structures, commonly referred to as random-coil (Bhowmick *et al.*, 2016; Das *et al.*, 2015). Thus, the Gaussian-chain model which was proven to successfully simulate the intramolecular distance distributions in polypeptides exposed to highly concentrated chaotropic agents (Zhou, 2001, 2002a), may be providing a baseline to capture the intricate tendency of intrinsically disordered proteins to adopt partially folded, non-canonical structures. A detailed structural propensity analysis of human alpha-synuclein clearly demonstrates the both the N- and C-terminal parts of the protein populate a non-canonical structures, which deviate from the 'random-coil' behavior (Tamiola and Mulder, 2012). Consequently, the pair-wise distance distributions of charged moieties in intrinsically disordered alpha-synuclein are expected not to follow behavior predicted by Gaussian-chain model implemented in **pepKalc.** Thus, our computations can identify sites where locally persistent interactions may shift a pK_a constant to a value beyond that expected for a featureless chain, and could hence provide novel structural information.

In the current implementation, **pepKalc** and all previous published methods, which relie on simplified Debye-Hückel electrostatics, do not account for the electrostatic effects caused by highly concentrated protein in solution. Protein–protein and protein–solvent interactions are known to cause a significant and detectable effect on the ion binding affinities (Linse *et al.*, 1995), as well as have a profound effect on the overall dielectric properties of their solutions (Bibi *et al.*, 2016; Moser *et al.*, 1966). Thus, in a current form, given the approximate nature of Debye-Hückel interactions in **pepKalc,** overestimation of intra-molecular charge–charge interaction energies is expected at very low ionic strength. The generality of the method described here remains, however, and trivially allows Equation 10 to be adapted to obtain more accurate calculations. In the future, rather than the Gaussian Chain model for unfolded proteins, explicit three-dimensional ensembles of structures could be used. In addition, the pH-dependent structural deformations that take place in the case of particular charge distributions (Das and Pappu, 2013) could be considered. These refinements would all help to model protein unfolded states in a more realistic manner, and thereby improve the accuracy of the prediction.

Finally, the method has the innate flexibility to consider any type of acid or base in the polypeptide sequence, such that non-natural amino acids and post-translational modifications can also be included. These extensions are also offered by **pepKalc.**

5 Conclusion

We have developed **pepKalc,** a robust simulation software for the comprehensive evaluation of protein electrostatics in disordered polypeptides. Our software completely removes the limitations of previously reported Monte-Carlo approaches in the computation of protein electrostatics, by using a hybrid methodology that effectively combines exact and mean-field calculations to rapidly obtain accurate results. Paired with a modern architecture GPU, **pepKalc** is capable of evaluating protonation behavior for a typical protein in seconds. Our approach can be combined with other polymer models, including explicit ensembles. Thus the deviations from the protonation behavior obtained in high accuracy **pepKalc** simulations might be helpful in structure validation or as ensemble restraints for intrinsically disordered proteins with variable levels of secondary structure propensity.

Acknowledgements

Authors thank Alison Lowndes and Carlo Ruiz, (NVIDIA Corporation) for facilitating collaboration and access to DGX-1 supercomputing node.

Funding

This work was supported in part by a VIDI Grant to F.A.A.M. from The Netherlands Organization for Scientific Research (NWO)—Grant no. 700.56.422.

Conflict of Interest: none declared.

References

Alexov,E. *et al.* (2011) Progress in the prediction of pKa values in proteins. *Proteins Struct. Funct. Bioinf.*, **79**, 3260–3275.

Bashford,D. and Karplus,M. (1990) pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, **29**, 10219–10225.

Bhowmick,A. *et al.* (2016) Finding our way in the dark proteome. *J. Am. Chem. Soc.*, **138**, 9730–9742.

Bibi,F. *et al.* (2016) A review: origins of the dielectric properties of proteins and potential development as bio-sensors. *Sensors*, **16**, 1232.

Bombarda,E. and Ullmann,G.M. (2010) pH-dependent pKa values in proteins: a theoretical analysis of protonation energies with practical consequences for enzymatic reactions. *J. Phys. Chem. B*, **114**, 1994–2003.

Cookson,M.R. (2005) The biochemistry of Parkinson's disease. *Annu. Rev. Biochem.*, **74**, 29–52.

Croke,R.L. *et al*. (2011) NMR determination of pKa values in α-synuclein. *Protein Sci*., **20**, 256–269.

Das,R.K. and Pappu,R.V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. USA*, **110**, 13392–13397.

Das,R.K. *et al*. (2015) Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struc. Biol*., **32**, 102–112.

Elcock,A.H. (1999) Realistic modeling of the denatured states of proteins allows accurate calculations of the pH dependence of protein stability. *J. Mol. Biol*., **294**, 1051–1062.

Gear,C.W. *et al*. (1967) The numerical integration of ordinary differential equations. *Math. Comput*., **21**, 146–146.

Geist,L. *et al*. (2013) Protonation-dependent conformational variability of intrinsically disordered proteins. *Protein Sci*., **22**, 1196–1205.

George,J.M. (2002) The synucleins. *Genome Biol*., **3**, 3002–3010.

Gilson,M.K. (1993) Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins Struct. Funct. Genet*., **15**, 266–282.

Hass,M.A. and Mulder,F.A.A. (2015) Contemporary NMR studies of protein electrostatics. *Annu. Rev. Biophys*., **44**, 53–75.

Hykes,S. (2018) Docker - Build, Ship, and Run Any App, Anywhere. San Francisco.

Jha,A.K. *et al*. (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA*, **102**, 13099–13104.

Kumar,S. and Nussinov,R. (2002) Close-range electrostatic interactions in proteins. *ChemBioChem*, **3**, 604–617.

Lindman,S. *et al*. (2006) Electrostatic contributions to residue-specific protonation equilibria and proton binding capacitance for a small protein. *Biochemistry*, **45**, 13993–14002.

Linse,S. *et al*. (1995) The effect of protein concentration on ion binding. *Proc. Natl. Acad. Sci. USA*, **92**, 4748–4752.

Moser,P. *et al*. (1966) Electric polarization in proteins. Dielectric dispersion and Kerr effect studies of isoionic bovine serum albumin. *J. Phys. Chem. A*, **70**, 744–756.

Nielsen,J.T. and Mulder,F.A.A. (2016) There is diversity in disorder-"In all Chaos there is a Cosmos, in all Disorder a Secret Order". *Front. Mol. Biosci*., **3**, 4.

Pierattelli,R. and Felli,I.C. (eds) (2015) *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*. Springer International Publishing, Switzerland.

Receveur-Brechot,V. and Durand,D. (2012) How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci*., **13**, 55–75.

Tamiola,K. and Mulder,F.A.A. (2012) Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. Trans*., **40**, 1014–1020.

Tanford,C. (1961) *Physical Chemistry of Macromolecules*. Wiley, New York.

Tanford,C. and Kirkwood,J.G. (1957) Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc*., **79**, 5333–5339.

van Rossum,G. (1995). Python tutorial. Technical report Centrum voor Wiskunde en Informatica (CWI) Amsterdam.

Wallerstein,J. *et al*. (2015) Site-specific protonation kinetics of acidic side chains in proteins determined by pH-dependent carboxyl 13C NMR relaxation. *J. Am. Chem. Soc*., **137**, 3093–3101.

Wood,S.J. *et al*. (1999) alpha-synuclein fibrillogenesis is nucleation-dependent. Implications for the pathogenesis of Parkinson's disease. *J. Biol. Chem*., **274**, 19509–19512.

Yang,A.S. *et al*. (1993) On the calculation of pKas in proteins. *Proteins Struct. Funct. Genet*., **15**, 252–265.

You,T.J. and Bashford,D. (1995) Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. *Biophys. J*., **69**, 1721–1733.

Zhou,H.X. (2001) Loops in proteins can be modeled as worm-like chains. *J. Phys. Chem. B*, **105**, 6763–6766.

Zhou,H.X. (2002a) Residual charge interactions in unfolded *Staphylococcal* nuclease can be explained by the Gaussian-Chain Model. *Biophys. J*., **83**, 2981–2986.

Zhou,H.X. (2002b) A Gaussian-chain model for treating residual charge–-charge interactions in the unfolded state of proteins. *Proc. Natl. Acad. Sci. USA*, **99**, 3569–3574.