



High pKa variability of cysteine residues in structural databases and the effect of H-bond contributions

[Yapısal veritabanlarındaki sistein kalıntılarında pKa değerinin yüksek oranda değişkenlik göstermesi ve H-bağı katkılarının etkisi]

Stefano M. Marino,
İnanç Soylu

Akdeniz University, Department of Agricultural
Biotechnology, Antalya

Correspondence Address
[Yazışma Adresi]

Stefano Maria Marino, Ph.D.

Department of Agricultural Biotechnology,
Akdeniz University, Türkiye
Phone: +90 242 3102518
E-mail: marinost@akdeniz.edu.tr

Registered: 05 November 2013; Accepted: 10 June 2014
[Kayıt Tarihi: 05 Kasım 2013; Kabul Tarihi: 10 Haziran 2014]

ABSTRACT

Objective: Our first objective was to extensively compare two most common empirical protein pKa predictors, propka1.0 (ppka1) and propka3.0 (ppka3); we have specifically compared them as tools for high-throughput analyses of structural datasets with a particular focus on the amino acid Cysteine (Cys); afterwards, our goal was to assess their performances with known instances of reactive Cys residues.

Methods: Structural datasets were downloaded from the PDB repository and pipelined to different pKa prediction software; results were parsed with in-house scripts, to extract relevant information, and then subjected to further analysis, including detailed output comparisons for different programs.

Results: With ppka1, H-bond contributions dictated the prediction of Cys pKa, particularly for exposed residues; this was not the case for the most recent version, ppka3. This feature of ppka1 fits with recent, independent studies reporting the critical role of H-bond network in the activation of reactive Cys residues; indeed, when tested in a benchmark for its ability to describe reactive Cys residues, ppka1 provided the best results, favorably comparing to other methods tested.

Conclusion: ppka1 can be an effective aid in redox bioinformatics as a tool for high-throughput Cys pKa predictions: it is extremely fast, yet capable of competitive performances, particularly apt to predict very reactive (e.g. nucleophilic, exposed) functional Cys residues. This work provides new insights on propka (in its different versions) predictions as well as substantial support to the critical role of H-bond and exposure in Cys activation.

Key Words: Cysteine, redox biology, bioinformatics, biochemistry

Conflict of Interest: Authors have no conflict of interest.

ÖZET

Amaç: Bu çalışmada, yaygın olarak kullanılan iki ampirik pKa tahmin edici program olan propka 1.0 (ppka1) ile propka 3.0 (ppka3)'ü karşılaştırılmıştır. İki program, sistein (Cys) amino asidi odaklı, yapısal veri tabanlarının yüksek hacimli analizlerinde kullanılan araçlar olarak test edilmiş, ve performanslarının, bilinen reaktif sistein kalıntılarına ilişkin bilgiler kullanılarak değerlendirilmesi hedeflenmiştir.

Metod: Yapısal veri setleri, PDB veri tabanından bilgisayara indirilmiş ve farklı pKa tahmin programların aktarılmıştır. Programların çıktılarında amaca yönelik hazırlanmış betikler kullanılarak ilgili bilgiler çıkarılmış ve elde edilen bilgiler daha sonra farklı programların çıktılarının karşılaştırıldığı detaylı analizlere konu edilmiştir.

Bulgular: ppka1 programında, Cys kalıntısına ilişkin pKa değerinin hesaplanırken; özellikle yüzeyde yer alan kalıntılarda, H-bağı katkıları önemli bir yer teşkil etmektedir. Son sürüm olan ppka3 için bu durum farklıdır. ppka1'nın bu özelliği H-bağı ağının reaktif, fonksiyonel Cys kalıntılarının aktivasyonundaki kritik rolünü gösteren, yakın zamanda yapılmış bağımsız çalışmalar ile örtüşmektedir. Gerçekten de reaktif Cys kalıntılarının tanımlayabilme yeteneği test edildiğinde, ppka1 programından diğer metotlara kıyasla daha başarılı sonuçlar elde edilmiştir.

Sonuç: Çalışmada ppka1 programının, redoks biyoinformatiğinde, yüksek hacimli veri setlerinin Cys kalıntılarında tahmininde etkin bir araç olarak kullanılabileceği gösterilmiştir. ppka1 özellikle reaktif (Örn. nükleofilik, yüzeyde yer alan) fonksiyonel Cys kalıntılarının hassas olarak tahmin edilmesinde, hızlı ve diğer programlarla rekabet edebilen performansa sahiptir. Bu çalışma, propka (ppka1, ppka3) tahminleri ile ilgili ilişkin önemli bilgiler ile beraber Cys aktivasyonunda H-bağının ve kalıntının yüzeyde bulunmasının kritik rolünü destekleyici bilgiler sunmaktadır.

Anahtar Kelimeler: Sistein, redoks biyoloji, biyoinformatik, biyokimya

Çıkar Çatışması: Yazarların çıkar çatışması yoktur.

Introduction

Cysteine (Cys) is a very versatile amino acid, often found at the functional sites of proteins, where it serves a variety of structural and functional roles [1-3], including structural stabilization (disulfide bridges, metal coordination), catalysis (e.g. active site nucleophile in thiol-based oxidoreductases) and post-translational modifications (e.g. reversible Cys modifications through reactive oxygen species, ROS, and reactive nitrogen species) [4,5]. Apart from its standard physiological roles, Cys is also a major target of oxidative stress: at higher ROS concentrations, non-functional, but solvent accessible, Cys residues may incur unwanted oxidation events [6-8]; in many cases, these oxidations are effectively counteracted through natural defensive mechanisms [9]; however, intense and/or persistent levels of ROS stress ultimately lead to protein damage (misfolding, cross-linkages, inactivation of functional residues) [10,11]. A critically important aspect of Cys reactivity is its acid dissociation constant (e.g. pKa of its thiol, i.e. Cys functional group). Most commonly, deprotonated Cys are better nucleophiles than their protonated counterparts, and tend to be significantly more reactive toward ROS and, in general, electrophilic substrates [12,13]. Thus, effective prediction protocols of Cys pKa are very important, particularly for studies aimed to detect and describe reactive Cys residues in proteins; for example, the identification of specific redox targets in the cell (e.g. substrates of thiol-oxidoreductases) or prediction of new hot spots of oxidative stress.

Different approaches have been applied to Cys pKa prediction: *a*) density functional theory (DFT) based predictors, deriving pKa estimation from natural population analysis on the atomic charge on Cys sulfur [14]; *b*) numerical solutions of the Poisson-Boltzmann (PB) equation, calculating the probability of protonation at different pH [15,16]; and *c*) empirical protocols, where energetic contributions provided by surrounding residues are estimated based on semi-empirical rules (i.e. empirically derived approximations, turned into theoretical assumptions) [17-19]. All these approaches present advantages and disadvantages [2]. DFT based calculations can be extremely precise [14]; PB methods are very informative (they provide detailed information on the titration range) [16]; empirical approaches are extremely fast, and much simpler to set up and handle [17-20]. However, both DFT and PB based methods implement complex theoretical protocols that severely limit their applicability to large scale automated calculations. Currently, empirical approaches are the only practical choice for large scale bioinformatics projects [2,21,22]: these methods are orders of magnitude faster, do not require extensive work to prepare input files, while at the same time maintaining good performances (as compared to major competitors) [23].

The most common empirical program is the method called propka. Three versions have been released: ver-

sion1.0 [17], 2.0 [18] and 3.0 [19]; all of them provide very fast computational performances (i.e. they are all suited for high-throughput bioinformatics tasks). The first two propka implementations differ only for the way protein ligands are treated; for our goals, (i.e. assessing databases of protein structures only) results from versions 1.0 (ppka1) and 2.0 (ppka2) are super imposable: we chose ppka1, as it is the original version. In turn, the latest version (propka 3.0, ppka3) implements a different parametrization of the empirical rules [19,20], and thus its results are different from ppka1 (and ppka2). In this work we specifically addressed this question: how different are the results from these two programs? And, if these differences are significant, what is the rationale behind diverging results? Among titratable residues (Asp, Glu, Tyr, Cys, His, Lys, Arg), we reserved specific attention to Cys, a major subject of our research interests. As to Cys, various independent studies have highlighted the critical importance of *(i)* the H-bond network (i.e. H-bond network more important than other contributors, like electrostatic perturbation, etc), and *(ii)* exposure, in determining its reactivity [2,14,22,24]: most reactive functional Cys tend to be solvent accessible (not surprisingly, as deeply buried sites are usually not accessible to reactants) and present tight H-bond networking. We carefully examined the behavior of different propka version in respect to these features, and against a large dataset of protein structures. For ppka1, we recorded a very tight relationship (i.e. more so than for any other amino acid) between Cys pKa and H-bond network contributions, particularly for solvent accessible Cys residues; the relationship was significantly attenuated in the most recent program, ppka3. This difference can have profound biological implications: reactive and functional Cys residue are often exposed and involved in H-bond interactions [2,24]; for them, ppka1 and ppka3 tend to present significantly different results, and thus should not be used interchangeably. To assess this claim, we tested them with a representative benchmark of well characterized, reactive Cys residues.

Materials and Methods

Structural datasets

The PDB repository [25] was used as the reference source of protein structures; to refine our dataset we applied the following filters to the search, using tools available for advanced searches implemented in the RCSB search engine [25]: *(i)* only crystallographic structures were considered; *(ii)* only structures with less than 30% amino acidic identity (to any other structure in the dataset) were kept; *(iii)* only protein with more than 100 and less than 1000 amino acids were considered. Proteins passing these filters were downloaded locally and, afterwards, checked for structural inconsistencies with the molecular modeling suite VegaZZ [26]. Considering its (very large) size, the complete list of proteins and residues, including the full details of the calculations reported in these results (and summed up

Table 1. Comparison of different propka implementations (all residues)

	ppka3 [pKa] [#]	ppka3 SD	ppka1 [pKa]	ppka1 SD	ΔpKa (ppka3 – ppka1)	p-value (t-test, n=1000)
CYS	10.953	1.325	9.151	1.985	1.802	2.5E-189
ASP	3.601	0.843	3.194	1.021	0.407	7.1E-21
GLU	4.395	0.724	4.142	0.89	0.253	1.2E-12
TYR	11.201	1.129	10.899	1.365	0.302	5.2E-08
HIS	5.872	1.060	5.916	1.719	-0.044	0.541
ARG	12.515	0.655	12.036	0.565	0.479	3.6E-55
LYS	10.485	0.614	10.333	0.461	0.152	1.3E-11

[#]legend: [pKa]=average pKa for each type, SD=Standard deviation; ΔpKa (ppka3 – ppka1) = ppka3 [pKa] - ppka1 [pKa]; t-test described in Methods; 1,000 data points (for each residue type) were randomly selected from the full dataset of 750 proteins.

Table 2. Comparison of different propka implementations (only exposed residues)

	ppka3 [pKa] [#]	ppka3 SD	ppka1 [pKa]	ppka1 SD	ΔpKa (ppka3 – ppka1)	p-value (t-test, n=1000)
CYS	9.545	0.685	8.220	1.352	1.325	4.1E-108
ASP	3.463	0.644	3.223	0.747	0.240	2.8E-09
GLU	4.323	0.581	4.187	0.604	0.136	1.8E-04
TYR	10.597	0.692	10.650	1.241	-0.053	0.515
HIS	6.292	0.488	6.656	0.406	-0.364	1.1E-32
ARG	12.594	0.499	12.064	0.320	0.530	4.2E-61
LYS	10.539	0.494	10.327	0.209	0.212	9.2E-19

[#] legend: [pKa]=average pKa for each type, SD=Standard deviation; ΔpKa (ppka3 – ppka1) = ppka3 [pKa] - ppka1 [pKa]; t-test described in Methods; 1,000 data points (for each type) were randomly selected from the full dataset (750 proteins).

in Table 1 and Table 2), are available as supporting information (Suppl. file 1, Suppl. file 2).

pKa predictions

For pKa predictions we used the empirical approach propka; we have downloaded and tested different versions: version 1.0 (ppka1, the original implementation, providing exactly the same results of version 2.0, ppka2, if results for ligands are not considered, as was the case in our calculations) and the latest implementation (version 3.0, ppka3). We compared in details the results of ppka1 and ppka3 (completed data are available as supporting information, Suppl. file 1, Suppl. file 2). The two versions of the propka program were then tested with an in-house conducted benchmark (Table 3), aimed to assess their performances against known cases of functional Cys residues. For a more complete picture, we compared ppka1 and ppka3 with another common pKa predictor, the PB based method H++ [27]. H++ calculations were run with default parameters (to normalize predictions, i.e. same set up for different proteins in Table 3). H++ was used only for benchmark purposes; as detailed in the Introduction, its application to high-throughput bioinformatics is hampered by the considerable computational load intrinsic to PB based calculations. In general, and with current (or near-future) hardware capabilities, similar approaches are

not a practical choice for high-throughput scans of structural databases.

Benchmark

For our benchmark, we selected a set of proteins from literature; because of limited experimental information, Cys is often avoided in protein pKa benchmarks [17,23]. To collect reference information we have looked for publications reporting experimentally derived Cys pKa; among them, we searched for cases where structural information was available (e.g. protein with a PDB entry). When literature is searched in this direction only a limited number of entries can be listed (in our case, ten proteins Table 3); as previously mentioned, this is mostly ascribable to the scarcity of reliable experimental data; however, it has to be noted that the size of our benchmark is comparable with previous works on Cys pKa predictions and benchmark [2,14,17,23], where the average number of proteins analyzed for Cys pKa (i.e. on average, seven) was below ten entries (the size of our set, Table 3). The structural references for all proteins in Table 3 are the following (gene name (*source*)/ PDB structural identifier): trxA (*H. sapiens*)/ 1XOB, grx3 (*H. sapiens*)/ 1ILB, pdi (*E. coli*)/ 4EKZ, grx1 (*E. coli*)/1JHB, trx2 (*R. capsulatus*)/ 2PPT, trxA (*S. aureus*)/ 2O89, dsba (*E. coli*)/ 1a24, grx1 (*S. cerevisiae*)/ 2JAC, grx1 (*S. scrofa*)/ 1KTE, TryX (*T. brucei*)/ 1O73.

Table 3. Benchmark

Gene	Source	Residue	"Exp"	"ppka3"	"ppka1"	"H++"#
trxA	<i>H. sapiens</i>	Cys51	4.8	7.7	4.2	10.2
grx3	<i>H. sapiens</i>	Cys22	3.6	8.6	5.5	7.7
pdi	<i>E. coli</i>	Cys11	5.5	7.6	4.3	7.6
grx1	<i>E. coli</i>	Cys32	7.1	9.2	4.8	6.1
dsba	<i>E. coli</i>	Cys30	3.5	10.7	9.5	10.1
trxA	<i>S. aureus</i>	Cys29	6.4	7.7	3.3	10.4
trx2	<i>R. capsulatus</i>	Cys73	5.2	8.3	5.9	10.4
grx1	<i>S. cerevisiae</i>	Cys26	4	7.9	4.7	8.3
grx1	<i>S. scrofa</i>	Cys22	4.9	8.2	4.1	8.8
TryX	<i>T. brucei</i>	Cys40	7.2	8.4	6.2	8.9

"Exp"= experimental data; other columns: results from different predictors (ppka1, ppka3, H++).

Statistical analysis

Considering the different amino acid abundance in proteins (e.g. titratable Cys residues, are typically much less than glutamic acids), statistics presented in Table 1 and 2 (i.e. summary of results for ppka1 and ppka3) were derived with the following procedure, aimed to normalize the data: 1,000 residues were employed for each titratable type; starting from the full dataset (40605 titratable sites, described before), 1000 data points for each type were randomly selected and further considered (i.e. exactly 1000 Asp, 1000 Arg, 1000 Tyr, 1000 Cys, 1000 Glu, 1000 His, 1000 Lys were compared). As such, the values for different amino acids can be better compared. The process of random selection was unbiased (i.e. not dependent on our choices and/or our supervision), made through automatic routines (i.e. shuffle functions, implemented in random class, python version 2.7). Statistical evaluations (average and deviations, linear correlations with Pearson's R coefficients) were performed with in-house python (version 2.7.4) scripts and with Microsoft Excel® 2010 (ANOVA, paired t-test, coefficients of determination). ANOVA (one way) test was run for data in Table 3, to confirm (null hypothesis not rejected; p-value=0.1462) that the groups have equal mean. This allowed further comparisons (e.g. how well each method's results could correlate with experimental data, Table 3), described in the results section (i.e. when discussing results of the benchmark).

Results and Discussion

Here we have employed the two most representative versions of the empirical pKa predictor propka, ppka1 (the original version) and ppka3 (the most recent version), for large scale, comparative analyses of their performances against protein structure datasets. For our calculations with ppka1 and ppka3, we built a dataset of proteins applying the following rules to the PDB repository: (i) only protein structural models derived from crystallographic studies were downloaded; (ii) to limit redundancy, only structures with less than 30% amino acidic identity (to

any other structure in the dataset) were kept; (iii) only protein with more than 100 and less than 1000 amino acids were considered. Among the PDB entries passing these filters (circa 13000, as of April 2013), 750 structures were randomly selected for further analysis, containing a total number of 40605 titratable sites. General statistics of the result of the application of ppka1 and ppka3, run in parallel and on the same dataset of 750 protein structures, are shown in Table 1. As expected, ppka1 and ppka3 provide different results. The differences are significant for all titratable types (t-test, $n=1000$; all p -values <0.001), except for His (i.e. p -value 0.541). To be noted: to derive the numerical values reported here, 1000 data points for each type were randomly selected (with automatic routines, to avoid internal bias, see Methods) and further considered; this was done to allow better comparisons between different amino acid types, which have different abundance (e.g. Lys, Glu and Asp are two to three times more frequent than Cys, His and Tyr). It should be further noted that 1000 points per amino acid, randomly selected through unbiased computational procedures, should represent a statistically robust dataset; we confirmed this claim by performing three independent runs of "random sampling" (i.e. three repetition of the random selection, each time independently randomizing the selection of the 1000 residues, for each amino acid type); these control calculations (available as supporting information, Suppl. file 3) conclusively confirmed the data in Table 1 (i.e. regardless of the randomly chosen set of residues, results in Table 1 are consistently valid), which thus can be safely commented and further discussed. A first important aspect to note is the "outstanding case" of Cys: by all means, these residues showed the largest difference between ppka3 and ppka1 predictions (on average $\Delta pK_a=1.802$, a large value, with a very significant meaning). Considering the numbers involved, ppka1 predicts Cys with pKa values much closer (than what estimated by ppka3) to physiological ranges (e.g. 6 to 8 pH units). This observation is especially important for exposed residues: in these

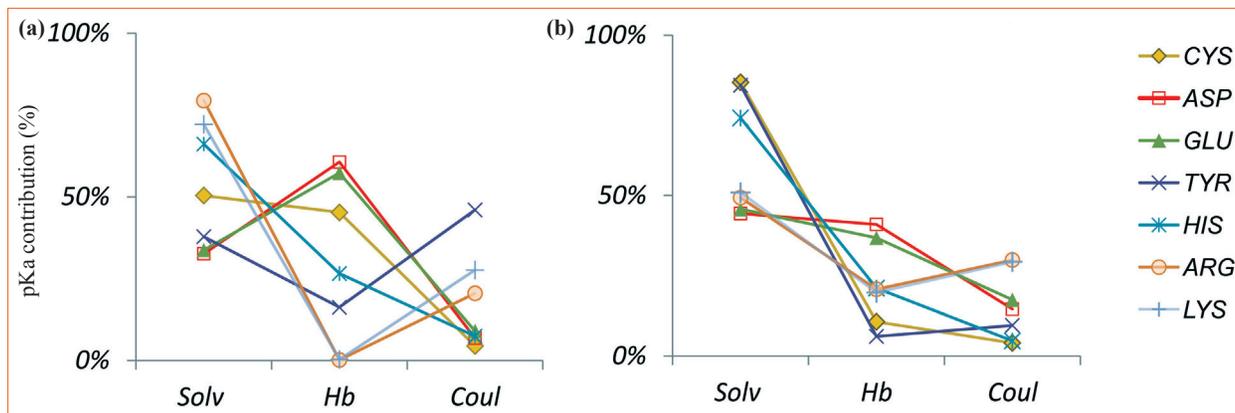


Figure 1. Relative weights distribution for ppka1 and ppka3. Each relative contribution refers to the weight of each component (Solvation effect, abbreviated as “Solv”; H-bond contribution, “Hb”; Coulombic interactions, “Coul”) on the final output (i.e. on the predicted pKa value), for each amino acid type. Relative weights are expressed in percentage (sum of weights=100%, for both panel A and B). Panel A (left) shows results relative to ppka; panel B (right) shows results relative to ppka3. The legend (reporting symbols employed, for both panels) is shown on the right side of the figure. A note: lines connecting dots in the graphs do not represent trends (or any interpolating function), but only serve illustration purposes (i.e. so that values for each amino acid are easier to distinguish).

positions, thiolates being generally more reactive than thiols, such differences may have deeper functional implications. Table 2 sums up the results obtained running the two programs against the same dataset, but parsing in only residues in solvent accessible areas of the protein. The results are similar to Table 1: (i) Cys is still the most variable residue, and (ii) predictions of the two programs diverge the most in the case of Cys (again, the difference is significant; t-test, $n=1000$, $p\text{-value}<0.001$). The difference between ppka1 and ppka3 is still the highest in the comparison ($\Delta pK_a = 1.325$), but in this context it has even more relevant implications: from a biological perspective, ppka1 results imply that, at neutral pH, a large fraction of exposed Cys population exist in its most reactive form, i.e. as a thiolate. This is not the case for ppka3; while in-depth biophysical considerations are beyond the scope of this work, here we report that such differences in the output of propka versions are (i) particularly significant, (ii) particularly large, and (iii) particularly (biologically) relevant in the case of Cys. From various points of views and especially in respect to this important amino acid, ppka1 and ppka3 behave as significantly different programs (rather than “simply” different implementations of the same theoretical protocol); the difference is not merely computational, but has important biological appendices: ppka1 estimations imply a much larger population of reactive Cys thiolates in molecular surfaces, than ppka3. This makes the choice of which version of propka to employ very important for Cys predictions, as it will heavily affect the “biological significance” of the results. Such a clear cut observation has not been reported before; its implications are of highly practical relevance.

We wanted to dig further into this direction, investigating the reasons behind such differences. As previously introduced, during our calculations we have parsed each output file, separately collecting all the available informa-

tion, in terms of energetic contributors (solvation effect, H-bond network, and electrostatics) to the prediction. A major distinction between ppka1 and ppka3 resides in the relative weights scheme; in Fig. 1 the relative weights (measured as % of the contribution to the output) for ppka1 and ppka3 are reported: as expected (i.e. results from Table 1), these differences regard all amino acids, with His being the most “untouched” type, i.e. ppka1 and ppka3 give a similar treatment to this amino acid. Also acidic residues are comparably evaluated, with H-bond seen by both methods as the biggest contributor. The two most diverging cases are Cys and Tyr; here solvation effects are more prominent in ppka3 than in ppka1, to the considerable detriment of H-bond network (Cys) and electrostatics (Tyr). Focusing on Cys (Fig. 2, and legend),

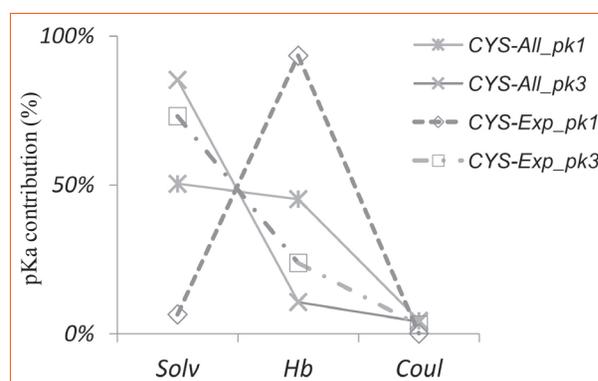


Figure 2. Relative weights distribution for Cys pKa predictions. The relative contributions (expressed in %; sum of weights=100%) for Cys are shown; abbreviations are as in Fig.1 and Table 1. Straight lines connect data-points for all-residue analysis; dotted lines, for exposed residues only. Squares and rhombi denote data from ppka1 (further abbreviated to pk1 in the Legend panel depicted in the figure), crosses and asterisks mark data from ppka3 (further abbreviated to pk3 in the legend panel). Like Fig.1, lines connecting dots in the graphs do not represent trends, but only serve illustration purposes.

the all-residues analysis (straight lines in Fig. 2) show that ppka1 weights much more (than ppka3) the H-bond contribution, while less the solvation effect; on average, the latter contributes an outstanding 85% in the decision process of ppka3; this value is downplayed to 50% in the case of ppka1. When only exposed Cys residues are considered, the difference between ppka1 and ppka3 becomes substantial; two major observations: (i) ppka1 weights the effect of exposure more than ppka3 (difference between square points and crosses in Fig. 2); (ii) ppka1 decision is driven by H-bond network, with solvation effect becoming secondary (if anything, the opposite scheme holds true for ppka3). Overall, ppka1 behave in a more “extreme” way (i.e. the effect of exposure on weights distribution is very pronounced). We have further investigated the details of the contribution of H-bond networks on exposed Cys pKa predictions: in Fig. 3, the correlation plot is shown, for ppka1 (left panel) and ppka3 (right panel). The difference between the two can be now visually appreciated: for ppka1 a strong linear correlation ($R^2=0.977$) is evident, indicating a very tight, direct effect of H-bond contribution to the final output (i.e. pKa); for ppka3, this tight relationship is lost, and the correlation is much weaker ($R^2=0.513$): ppka3 significantly downweights the importance of the H-bond network. Altogether, ppka1 and ppka3 differ the most in the case of solvent accessible Cys residues, due to a considerably diverging weight distribution scheme. As introduced before, recent DFT-based studies individuated H-bond network as the most critical determinant of Cys pKa in reactive Cys residues [14,24]; considering the high level of theoretical detail and accurate descriptions of the chemical system allowed by DFT, these results are particularly noteworthy. By overweighting H-bond contribution, ppka1 could stand as (significantly more so than ppka3) an effective

tool for the analysis of reactive Cys pKa, particularly in high-throughput analyses (where DFT cannot possibly apply). We have tested this claim with an in-house, manually curated benchmark: ppka1 and ppka3 were run in parallel with a test case of well known and characterized proteins, each containing a reactive, functional Cys residue (further details provided in the Methods section). As an additional comparative measure, we run another common pKa predictor, H++, a representative of the PB based methods [16], described in the introduction. This benchmark is comparable in size and objectives to previous literature reports (we refer in the Methods section for more details). Ppka1 provided the most accurate results: the average difference, between experimental and calculated values were 1.8 pH units, favorably comparing to a corresponding value of 3.2 pH units for ppka3 and 3.8 units for H++. Furthermore, we have tested which predictive method could correlate best with the experimental data; particularly we tested each set of predictions (i.e. “ppka3”, “ppka1”, and “H++” columns in Table 3) for independence with the experimental data (column “Exp” in Table 3). With the null hypothesis being that values in different columns were independent ($\alpha=0.05$) results showed that (i) ppka3 did not sufficiently describe reactive Cys (data confirmed the null hypothesis, $p\text{-value}=0.000346$, $p\text{-value}<\alpha$); this results can be taken as a measure of poor correlation between ppka3 predictions and experimental data; (ii) ppka1 performed considerably better, as the data (column “ppka1” and column “Exp”) were found to be not independent (null hypothesis was not confirmed, $p\text{-value}=0.968454$, $p\text{-value}>\alpha$); (iii) like ppka3, H++ did not sufficiently describe reactive Cys (data confirmed the null hypothesis, $p\text{-value}=0.000777$, $p\text{-value}<\alpha$). Then, we run the same comparisons among different predictors: first, in support of previous results, ppka3 and

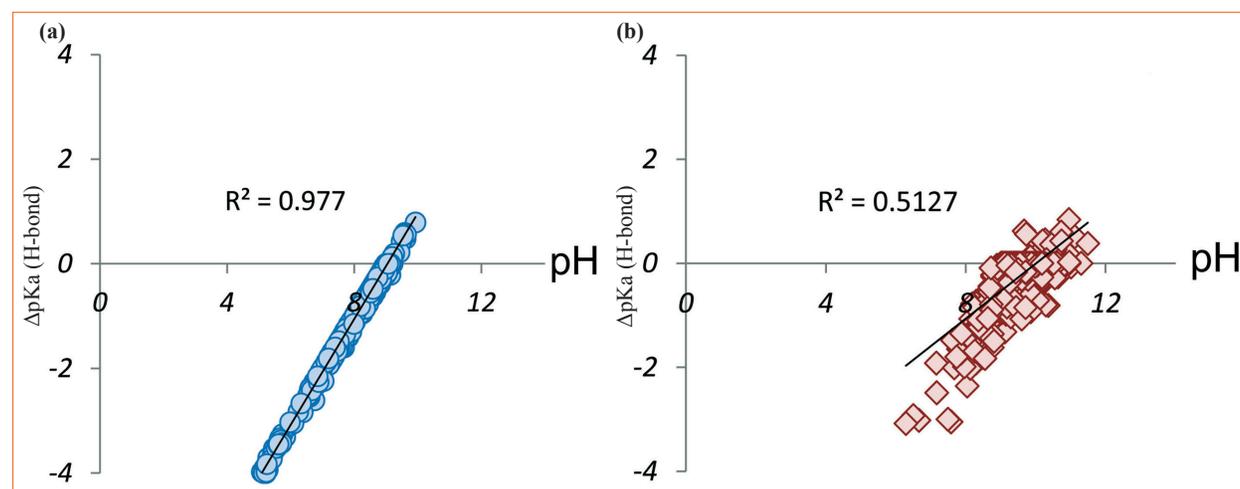


Figure 3. Correlation between H-bond energy contribution to pKa for exposed Cys residues. The figure shows the correspondence between predicted pKa (output) and H-bond network contributions for exposed Cys residues, in the case of ppka1 (plot on the left, circles in panel A) and ppka3 (plot on the right, rhombi in panel B). Correlation coefficients (R^2) are reported in the plots; in panel A (ppka1 predictions) the linear correlation is evident, denoting an outstanding prominence of H-bond interactions in determining predictions for exposed Cys residues. The correlation is much weaker for ppka3.

ppka1 were found to be not correlated (i.e. analysis on “propka1” and “propka3” columns in Table 3, confirmed the null hypothesis of independence, $p\text{-value}=4.4 \text{ E-}06$, $p\text{-value}<\alpha$). Interestingly, the results of ppka3 were closer to those obtained with H++ than to those with ppka1; indeed, if H++ and ppka1 predictions were compared, the null hypothesis (“predicted values are independent”) is confirmed, $p\text{-value}=0.000828$, $p\text{-value}<\alpha$, i.e. ppka1 and H++ output significantly different results; in turn, if H++ is compared with ppka3, the same hypothesis is rejected ($p\text{-value}=0.526297$, $p\text{-value}>\alpha$), i.e. H++ and ppka3 methods provide non significantly different results. All in all, at least in our test case (Table 3), ppka3 predictions are closer to those of a PB based approach, than to ppka1. Our results provide substantial support to the view that propka original implementations (e.g. ppka1) appear best suited to high-throughput analysis of reactive (e.g. particularly if exposed and nucleophilic, Table 3) functional Cys residues in structural databases.

Conclusion

In this work we have presented a large scale study on the performances of one of the most common methods for pKa prediction, propka; we tested two different implementations, the original program, ppka1, and the latest implementation, ppka3. A large, unbiased set of titratable residues derived from the PDB repository was employed for the calculations. We could register important differences among the two programs; these were particularly evident in the case of Cys. The major contribution behind this behavior was the different scheme of weights distribution; prominently, a dramatically different evaluation of the H-bond network contributions, which is highly regarded in ppka1 and downplayed in ppka3, especially for exposed Cys. Recent and independent studies on Cys reactivity [2,14,22,24] showed that solvent accessibility and involvement in H-bond networks, are the biggest determinant of Cys reactivity. Accordingly, our data suggest that ppka1, by providing particular attention and weight to H-bond contributions and exposure, appears well equipped to describe most reactive (e.g. exposed and with relatively low pKa, in particular) functional Cys. We have supported this conclusion with a benchmark, where ppka1 obtained the best results, and proved capable of better descriptions of a set of well known reactive Cys residues. Altogether, this study provides novel data and original insights into Cys pKa prediction; besides, it can be of practical interest for researches involved in Cys bioinformatics, and more broadly, in redox biology and thiol mediated regulation of protein function.

Acknowledgements

Conceived and designed the experiments: SMM. Performed the experiments: SMM and IS. Analyzed the data: SMM. Wrote the paper: SMM and IS.

This work was supported by a grant from TUBITAK (113Z524) to SMM.

Conflict of Interest

There are no conflicts of interest among the authors.

References

- [1] Beeby M, O'Connor BD, Ryttersgaard C, Boutz DR, Perry LJ, *et al.* The genomics of disulfide bonding and protein stabilization in thermophiles. *PLoS Biol* 2005; 3(9):309.
- [2] Marino SM, Gladyshev VN. Analysis and functional prediction of reactive cysteine residues. *J Biol Chem* 2012; 287(7):4419-25.
- [3] Cremers CM, Jakob U. Oxidant sensing by reversible disulfide bond formation. *J Biol Chem* 2013; 288(37):26489-96.
- [4] Miki H, Funato Y. Regulation of intracellular signalling through cysteine oxidation by reactive oxygen species. *J Biochem* 2012; 151(3):255-61.
- [5] Marino SM, Gladyshev VN. Structural analysis of cysteine S-nitrosylation: a modified acid-based motif and the emerging role of trans-nitrosylation. *J Mol Biol* 2010; 395(4):844-59.
- [6] Leonard SE, Carroll KS. Chemical ‘omics’ approaches for understanding protein cysteine oxidation in biology. *Curr Opin Chem Biol* 2011; 15(1):88-102.
- [7] Poole LB, Karplus PA, Claiborne A. Protein sulfenic acids in redox signaling. *Annu Rev Pharmacol Toxicol* 2004; 44:325-47.
- [8] Shenton D, Grant CM. Protein S-thiolation targets glycolysis and protein synthesis in response to oxidative stress in the yeast *Saccharomyces cerevisiae*. *Biochem J* 2003; 374(Pt 2):513-9.
- [9] D’Aur eaux B, Toledano MB. ROS as signalling molecules: mechanisms that generate specificity in ROS homeostasis. *Nat Rev Mol Cell Biol* 2007; 8(10):813-24.
- [10] Held JM, Gibson BW. Regulatory control or oxidative damage? Proteomic approaches to interrogate the role of cysteine oxidation status in biological processes. *Mol Cell Proteomics* 2012; 11(4):R111.013037.
- [11] Nakamura T, Cho DH, Lipton SA. Redox regulation of protein misfolding, mitochondrial dysfunction, synaptic damage, and cell death in neurodegenerative diseases. *Exp Neurol* 2012; 238(1):12-21.
- [12] Winterbourn CC, Metodiewa D. Reactivity of biologically important thiol compounds with superoxide and hydrogen peroxide. *Free Radic Biol Med* 1999; 27(3-4):322-8.
- [13] Winterbourn CC, Hampton MB. Thiol chemistry and specificity in redox signaling. *Free Radic Biol Med* 2008; 45(5):549-61.
- [14] Roos G, Foloppe N, Van Laer K, Wyns L, Nilsson L, *et al.* How thioredoxin dissociates its mixed disulfide. *PLoS Comput Biol* 2009; 5(8):1000461.
- [15] Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, *et al.* Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics Program. *Comp Phys Comm* 1995; 91:57-95.
- [16] Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, *et al.* H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 2005; 33(Web Server issue):W368-71.
- [17] Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* 2005; 61(4):704-21.
- [18] Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* 2008; 73(3):765-83.
- [19] Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions. *J Chem Theory Comput* 2011; 7(2):525-37.

- [20] Sondergaard CR, Olsson MHM, Rostkowski M, Jensen JH. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *J Chem Theory Comput* 2011; 7:2284-95.
- [21] Sanchez R, Riddle M, Woo J, Momand J. Prediction of reversibly oxidized protein cysteine thiols using protein structure properties. *Protein Sci* 2008; 17(3):473-81.
- [22] Marino SM, Gladyshev VN. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *J Mol Biol* 2010; 404(5):902-16.
- [23] Davies MN, Toseland CP, Moss DS, Flower DR. Benchmarking pK(a) prediction. *BMC Biochem* 2006; 7:18.
- [24] Roos G, Foloppe N, Messens J. Understanding the pK(a) of redox cysteines: the key role of hydrogen bonding. *Antioxid Redox Signal* 2013; 18(1):94-127.
- [25] Research Collaboratory for Structural Bioinformatics. Protein Data Base. <http://www.rcsb.org/pdb> (Last accessed: November 2013).
- [26] Drug Design Laboratory. VegaZZ software. <http://nova.colombo58.unimi.it/cms> (Last accessed: November 2013).
- [27] Virginia Polytechnic Institute and State University. H++ program. <http://biophysics.cs.vt.edu> (Last accessed: November 2013).