Nucleic Acids Research, 2006, Vol. 34, Database issue ***D199–D203***
doi:10.1093/nar/gkj035

PPD v1.0—an integrated, web-accessible database of experimentally determined protein pK_a values

Christopher P. Toseland, Helen McSparron, Matthew N. Davies and Darren R. Flower*

Edward Jenner Institute for Vaccine Research, Compton, Berkshire, RG20 7NN, UK

Received June 17, 2005; Revised September 15, 2005; Accepted September 26, 2005

ABSTRACT

The Protein pK_a Database (PPD) v1.0 provides a compendium of protein residue-specific ionization equilibria (pK_a values), as collated from the primary literature, in the form of a web-accessible postgreSQL relational database. Ionizable residues play key roles in the molecular mechanisms that underlie many biological phenomena, including protein folding and enzyme catalysis. The PPD serves as a general protein pK_a archive and as a source of data that allows for the development and improvement of pK_a prediction systems. The database is accessed through an HTML interface, which offers two fast, efficient search methods: an amino acid-based query and a Basic Local Alignment Search Tool search. Entries also give details of experimental techniques and links to other key databases, such as National Center for Biotechnology Information and the Protein Data Bank, providing the user with considerable background information. The database can be found at the following URL: http://www.jenner.ac.uk/PPD.

INTRODUCTION

A significant proportion of chemical reactions involving proteins are mediated through electrostatic interactions of their ionizable residues (1). Such residues greatly influence the conformation of a protein and therefore its function (2,3), as demonstrated by their folding mechanisms (4–6), enzyme catalysis and protein–protein interactions (7). With respect to enzyme catalysis, residues can act as proton donors and acceptors within the catalytic site and help stabilize transition states, with a concomitant influence on the rate of reaction (8,9).

The dissociation constant (K_a) is a measure of the acidity of a compound, i.e. its ability to donate a proton. K_a values range widely from 10^{10} for the strongest acids, such as sulphuric, to 10^{-50} for the weakest, such as methane. Therefore a negative logarithmic scale is usually applied (p$K_a = -\log_{10} K_a$), whereby K_a values for sulphuric acid and methane would become pK_a values of -10 and 50, respectively. Generally, more negative pK_a values correspond to stronger acids. The pK_a values of individual amino acid residues in proteins are determined by the ionization of their side-chain groups. For the 20 natural amino acids, pK_a values range from 4.0 for the side-chain carboxyl of aspartate to 12.0 for the side-chain guanididium group of arginine. Main-chain groups are not ionizable, although two additional ionizable groups exist at the N- and C-termini. Residues within proteins have pK_a values that are moderated by their micro-environments, the nature of their near neighbours, the extent of hydrogen bonding and so on and can take on a range of values different from that of a model residue.

NMR spectroscopy is the most widely used method for determining the pK_a values of individual residues, with an accuracy of $\sim$0.1 pH units. Although many NMR methods are available, most entries in the Protein pK_a Database (PPD) are derived using ^{1}H, ^{13}C and ^{15}N experiments. Inaccuracies in NMR experiments stem from the range of pH values tested, variations in ionic strength and the reversibility of the titration (10). In light of this, new combination methods are being used based on NMR spectroscopy coupled with site-directed mutagenesis, which leads to more accurate pK_a values (10,11).

The functional importance of ionizable residues has led to numerous attempts to predict individual residue-specific pK_a values (12–16). pK_a values are usually calculated from 3D structures using the Poisson–Boltzmann equation. However, variations occur between calculated and experimentally measured pK_a values (13). Molecular dynamic simulations have also been used for such predictions, although this only gives rise to a marginal increase in accuracy (17).

As only a small handful of reviews have attempted to compile residue-specific protein pK_a values (10,18,19), it was decided to develop a database that would serve as a standard compendium against which to compare new experimental or theoretical results. The PPD v1.0 contains >1400 amino acid pK_a values, sourced from experimental data. Cross-references to several external databases—the Protein Data Bank (PDB) (20), the Enzyme Nomenclature and Classification database

*To whom correspondence should be addressed. Tel: +44 1635 577954; Fax: +44 1635 577901 577908; Email: darren.flower@jenner.ac.uk

(21) and the National Center for Biotechnology Information (NCBI) Entrez-Protein—have also been incorporated into the database.

DATABASE DEVELOPMENT

PPD v1.0 has been implemented using a postgreSQL relational database, which provides an appropriate infrastructure for all foreseeable future developments of the archive. The data were initially compiled in a Microsoft ACCESS database after exhaustive searching of the primary literature, which included using keyword searches of the NCBI PubMed database (http://www.ncbi.nlm.nih.gov/pubmed). The postgreSQL database is structured into seven normalized tables, populated from a flat-file export of the ACCESS database using PERL scripts integrated with SQL. As data are continually accumulating, archiving data is an on-going process: automatic, periodic updates will be made to the postgreSQL database.

The PPD user interface is provided by a series of HTML pages. There are two searchable forms available within the PPD site. One offers either a broad or focussed PPD search. The other searches PPD using Basic Local Alignment Search Tool (BLAST). These forms target either a PERL/SQL script or a CGI script which in turn queries the database. The bespoke search engine facilitates fast, efficient and flexible data retrieval (Searching the Database). PPD is freely available on the world wide web (http://www.jenner.ac.uk/PPD).

DATABASE CONTENT

The data within PPD was sourced from the primary literature to give >1400 entries, containing pK_a values for >160 proteins (Table 1). The database contains pK_a values for amino acid side-chains, as well as the N- and C-termini. Data are archived for all amino acid residues, with the exception of methionine. However most entries focus on glutamate, lysine, histidine and aspartate, which together account for >75% of the data. As these four are all key ionizable residues, the apparent bias is not driven by our selection, but by the available experimental data. Very little data are currently available for arginine: its pK_a value ($\sim$12) essentially precludes measurement by titration as proteins will denature at such a high basic pH.

Cross-references to key external databases are also included. These provide links to the protein sequence, using NCBI Entrez-Protein, and any relevant protein structure in the PDB (20). If applicable, the enzyme classification is also given, with links to the Enzyme Nomenclature and Classification Database, developed in line with the International Union of Biochemistry and Molecular Biology (21), providing details of the enzyme reactions. In addition, a link is given to the original literature reference via the NCBI PubMed journals database. These links provide key background knowledge associated with each archived protein. A full description of the database fields is given in Table 2.

The ability to carry out accurate predictions of pK_a values depends on having access to a high quality source of data; a principal aim of PPD is to provide such a source. Only experimentally determined pK_a values are cited in PPD; predicted pK_a values are not included. The quality of data contained in PPD v1.0 is largely dependent upon the accuracy of each experimental determination, thus it contains only values from certain selected techniques: NMR spectroscopy, Raman Difference spectroscopy and UV spectroscopy.

Protein pK_a values are dependent on both intrinsic and extrinsic factors. Intrinsic factors include invariant properties of the protein investigated, such as sequence and structure. Extrinsic factors include the experimental conditions used, such as the temperature, the range of pH tested, protein concentrations as well as the experimental method. Thus we attempt to record all relevant experimental conditions when available. As logistic considerations preclude us from undertaking independent verification of the data, we are obliged to trust the values reported in the literature. It should be noted that the phenomenon of cooperative deprotonation can create circumstances under which pK_a values can not be used as a parameter that describes the ionization behaviour of the corresponding group (22–24).

SEARCHING THE DATABASE

Two methods to search PPD are available: an amino acid query-based interface (Figure 1) and a BLAST (25) interface. The implementation of a bespoke search system allows the user to perform extensive or focussed searches from a single user interface. The simplest search, using the amino acid query interface, would specify one amino acid residue only.

Table 1. Database summary

Database entries	1401				
Proteins					
Total	163				
PDB structures	146				
Sequences	115				
Enzymes	49				
Experiments					
Technique	^{13}C*	^{1}H*	^{15}N*	2D*	RS
Entries	235	780	46	112	56
Journals	189				

RS = Raman Difference Spectroscopy and * = NMR spectroscopy.

Table 2. Content of the database entries

Entry field	Description
Protein	States the relevant protein and provides a link to NCBI Entrez-Protein sequence
PDB	States the proteins PDB identification and provides a link to the structure
EC	The Enzymes Commissions identification and provides a link to the external database
Species	Species in which the protein is found
Protein description	Gives the basic function of the protein
Amino acid	The amino acid to which the pK_a refers
Residue	The residue number to which the pK_a refers
pK_a	pK_a value for the corresponding residue
Method	Experiment techniques used to obtain data, e.g. NMR
Temperature	Temperature at which the experiment was carried out
pH	Range or fixed pH at which the experiment were carried out
Conditions	Concentrations of substances used in the experiment
Unit intervals	Intervals at which recordings were taken (pH units)
Reference	Full literature reference with link to the PubMed database

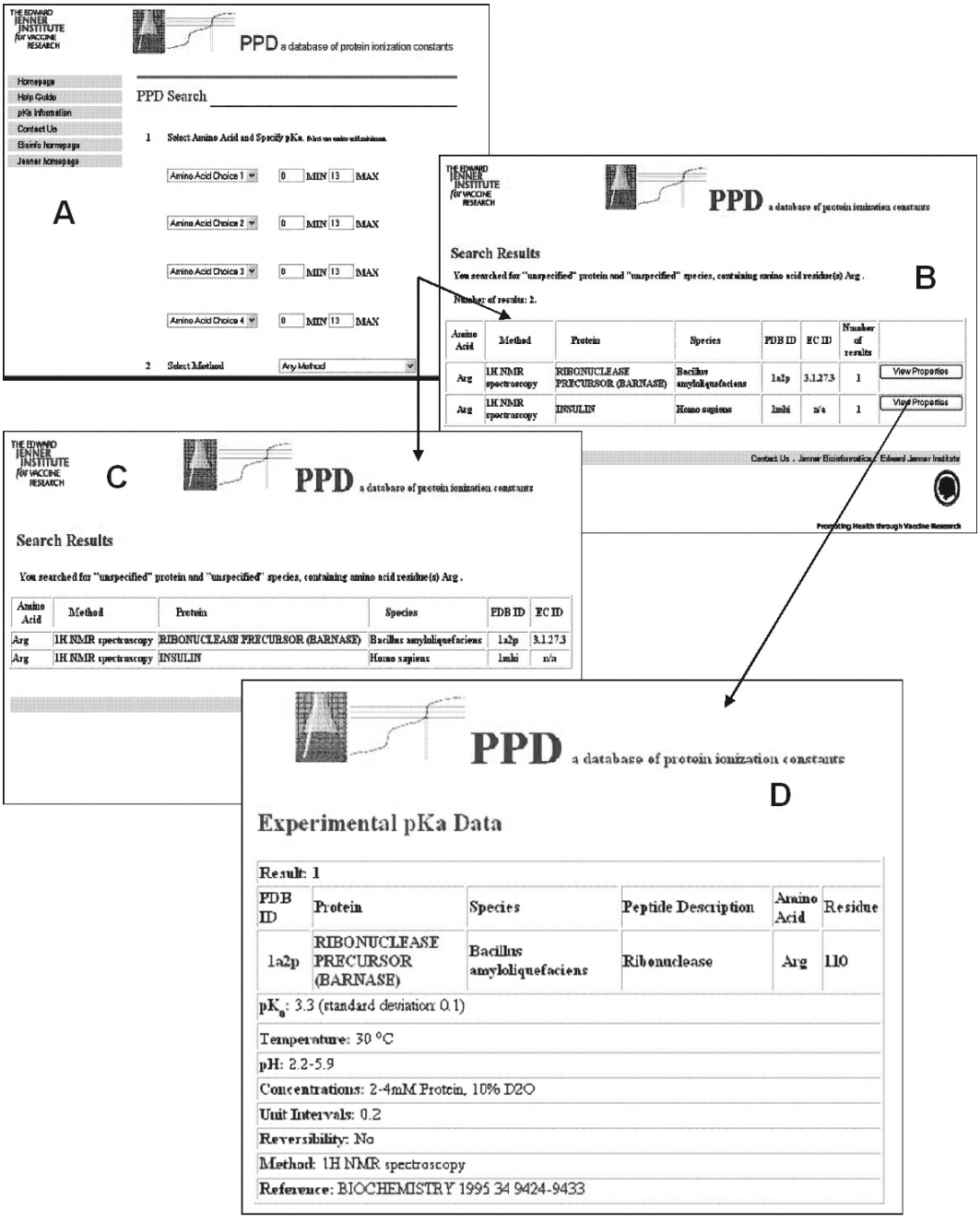

Figure 1. Overview of the amino acid query search. The amino acid nominations are entered in (**A**). (**B**) shows the default result presentation, from which the pK_a data (**D**) for the specified residues can be accessed. (**C**) shows the alternative presentation, with the display of proteins containing the nominated amino acid(s).

A complex search would accommodate up to four amino acids and pK_a ranges, along with experimental method, protein name and species. The search engine allows the choice of how results are presented. The default option returns amino acids and their associated properties (Figure 1B); while the second option returns proteins which contain the specified amino acids (Figure 1C).

The alternative search interface is based on BLAST (25). A local database of protein sequences found in PPD was compiled from SwissProt (26) and an additional postgreSQL table was created to hold this data. The local database is searched using the NCBI BLASTP and BLASTX programs (25), allowing input of either protein or nucleotide sequences. The HTML front-end connects to a web server-based PL/CGI script which interacts with the BLASTP or BLASTX programs. The output contains links to PPD entries, which are created using SwissProt (26) accession codes.

FUTURE WORK

There is an obvious need to extend the number of entries through continuous addition of data from new, and

newly-identified, publications. The database also needs to be maintained, ensuring links to external databases remain current. Initially, as with all databases, random errors will occur owing to human error during data acquisition or will be extant within the original experimental data. The database will be assessed for errors and inconsistencies, thus maintaining, as far as possible, the overall veracity of our data. As mentioned, we have tried to maintain a high degree of accuracy, through rigorous data selection; however, user feedback will foment improvements. Moreover, feedback focussing on the search interfaces and the general infrastructure will allow us to develop appropriately both the database and its interface in an efficient and ergonomic manner.

DISCUSSION AND CONCLUSIONS

The PPD is a unique compilation of protein pK_a values sourced from experimental data only. PPD is novel: no database of its kind currently exists. Compared with other post-genomic databases, the size of PPD is limited, but this reflects its highly focused nature: the burgeoning of such focussed databases is a continuing trend in modern bioinformatics (27,28). The relatively modest size of the database will increase as new data is published.

Access to PPD data is given through an interface available via the world wide web and includes both a BLAST search and an amino acid query search system. The BLAST search, which is linked to pK_a entries and external databases, allows PPD to be a cohesive and integrated source of protein information. PPD facilitates data-driven *in silico* prediction methods addressing the relationship between ionizable groups and protein function, be that protein–protein interaction, protein folding or enzyme catalysis.

A brief summary of pK_a data for each amino acid is shown in Table 3, which also includes both the mean and SD of the corresponding measured pK_a values. From the PPD data, we have shown the distribution of pK_a values for the six most frequent residues: glutamic acid, lysine, tyrosine, aspartic acid, histidine and cysteine (Figure 2). Certain residues (aspartate, glutamate, lysine and histidine) have pK_a values which show relatively narrow distributions, while other residues (cysteine and tyrosine) show a wider dispersion of values; however, this may only be a reflection of the amount of data available for these residues. While it is clear that mean values approximate closely model values, the corresponding SDs are high, reflecting the wide distribution of ionization states in

Table 3. pK_a data associated with each amino acid

Amino acids								
Residue	Asp	Cys	Glu	His	Lys	Tyr	N-terminus	C-terminus
Number of entries	282	25	297	404	207	65	26	38
Mean pK_a	3.6	6.87	4.29	6.33	10.45	9.61	8.71	3.19
SD	1.43	2.61	1.05	1.35	1.19	2.16	1.49	0.76

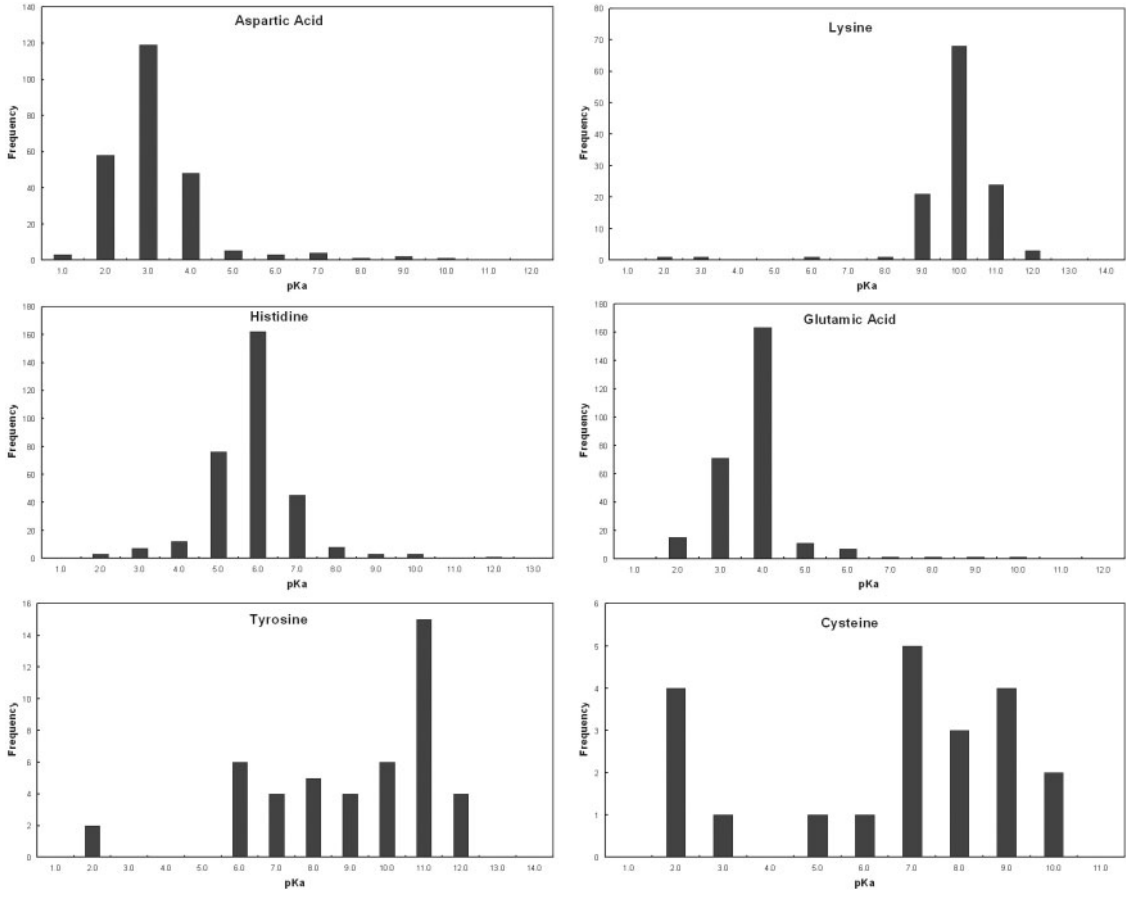

Figure 2. Distribution pattern of pK_a values. Each column represents a count of pK_a values for the specified amino acid and pK_a.

actual proteins. Aspartate, for example, has a mean pK_a of 3.6 versus a model value of 4.0, yet the SD is 1.4. As the data for each residue increases, trends in residue-specific pK_a data will become more evident and more certain.

In recent years, there has been an impetus to accumulate data on all scales from the atomic to the genomic; this has led to a rapid increase in the number of databases. Databases are increasingly forming the backbone of science in general and post-genomic biology in particular. PPD v1.0 was developed to provide an easily accessible compilation of protein pK_a values. Despite the small size of PPD, the data it contains has utility throughout many different disciplines and, we may hope, the database will grow, through time, into a comprehensive protein pK_a resource.

ACKNOWLEDGEMENTS

We should like to thank Andrew Worth for his technical assistance and Martin Blythe for programming advice. The Edward Jenner Institute for Vaccine Research wishes to thank its sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the UK Department of Health. Funding to pay the Open Access publication charges for this article was provided by the sponsors of the EJIVR.

Conflict of interest statement. None declared.

REFERENCES

1. Honig,B. and Nicholls,A. (1995) Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
2. Warshel,A. (1978) Energetics of enzyme catalysis. *Proc. Natl Acad. Sci. USA*, **75**, 5250–5254.
3. Antosiewicz,J., McCammon,J.A. and Gilson,M.K. (1994) Prediction of pH-dependent properties of proteins. *J. Mol. Biol.*, **238**, 415–436.
4. Lambeir,A.M., Backmann,J., Ruiz-Sanz,J., Filimonov,V., Nielsen,J.E., Kursula,I., Norledge,B.V. and Wierenga,R.K. (2000) The ionization of a buried glutamic acid is thermodynamically linked to the stability of Leishmania mexicana triose phosphate isomerase. *Eur. J. Biochem.*, **267**, 2516–2524.
5. Horng,J.C., Cho,J.H. and Raleigh,D.P. (2005) Analysis of the pH-dependent folding and stability of histidine point mutants allows characterization of the denatured state and transition state for protein folding. *J. Mol. Biol.*, **345**, 163–173.
6. Jamin,M., Geierstanger,B. and Baldwin,R.L. (2001) The pK_a of His-24 in the folding transition state of apomyoglobin. *Proc. Natl Acad. Sci. USA*, **98**, 6127–6131.
7. Norel,R., Sheinerman,F., Petrey,D. and Honig,B. (2001) Electrostatic contributions to protein–protein interactions: fast energetic filters for docking and their physical basis. *Protein Sci.*, **10**, 2147–2161.
8. Nielsen,J.E. and McCammon,J.A. (2003) Calculating pK_a values in enzyme active sites. *Protein Sci.*, **12**, 1894–1901.
9. Gerratana,B., Cleland,W.W. and Frey,P.A. (2001) Mechanistic roles of Thr134, Tyr160, and Lys 164 in the reaction catalyzed by dTDP-glucose 4, 6-dehydratase. *Biochemistry*, **40**, 9187–9195.
10. Forsyth,W.R., Antosiewicz,J.M. and Robertson,A.D. (2002) Empirical relationships between protein structure and carboxyl pK_a values in proteins. *Proteins*, **48**, 388–403.
11. Forsyth,W.R. and Robertson,A.D. (2000) Insensitivity of perturbed carboxyl pK(a) values in the ovomucoid third domain to charge replacement at a neighboring residue. *Biochemistry*, **39**, 8067–8072.
12. Warshel,A. (1981) Calculations of enzymatic reactions: calculations of pK_a, proton transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry*, **20**, 3167–3177.
13. Antosiewicz,J., McCammon,J.A. and Gilson,M.K. (1996) The determinants of pK_as in proteins. *Biochemistry*, **35**, 7819–7833.
14. Gogliettino,M.A., Tanfani,F., Scire,A., Ursby,T., Adinolfi,B.S., Cacciamani,T. and De Vendittis,E. (2004) The role of Tyr41 and His155 in the functional properties of superoxide dismutase from the archaeon *Sulfolobus solfataricus*. *Biochemistry*, **43**, 2199–2208.
15. Georgescu,R.E., Alexov,E.G. and Gunner,M.R. (2002) Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophys. J.*, **83**, 1731–1748.
16. Warwicker,J. (2004) Improved pK_a calculations through flexibility based sampling of a water-dominated interaction scheme. *Protein Sci.*, **13**, 2793–2805.
17. Bashford,D. and Gerwert,K. (1992) Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. *J. Mol. Biol.*, **224**, 473–486.
18. Edgcomb,S.P. and Murphy,K.P. (2002) Variability in the pK_a of Histidine side-chains correlates with burial within proteins. *Proteins*, **49**, 1–6.
19. Nielsen,J.E. and McCammon,J.A. (2003) On the evaluation and optimization of protein x-ray structures for pK_a calculations. *Protein Sci.*, **12**, 313–326.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. IUBMB, *Enzyme Nomenclature 1992*. Academic Press, San Diego.
22. Spitzner,N., Lohr,F., Pfeiffer,S., Koumanov,A., Karshikoff,A. and Ruterjans,H. (2001) Ionization properties of titratable groups in ribonuclease T1. I. pK_a values in the native state determined by two-dimensional heteronuclear NMR spectroscopy. *Eur. Biophys. J.*, **30**, 186–197.
23. Koumanov,A., Spitzner,N., Rüterjans,H. and Karshikoff,A. (2001) Ionization properties of titratable groups in ribonuclease T1. II. Electrostatic analysis. *Eur. Biophys. J.*, **30**, 198–206.
24. Koumanov,A., Rüterjans,H. and Karshikoff,A. (2002) Continuum electrostatic analysis of irregular ionization and proton allocation in proteins. *Proteins*, **46**, 85–96.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
27. Chalk,A.M., Warfinge,R.E., Georgii-Hemming,P. and Sonnhammer,E.L.L. (2005) siRNAdb: a database of siRNA sequences. *Nucleic Acids Res.*, **33**, D131–D134.
28. Mika,S. and Rost,B. (2005) NMPdb: database of nuclear matrix proteins. *Nucleic Acids Res.*, **33**, D160–D163.