Global Journal of Biotechnology & Biochemistry 1 (1): 28-34, 2006
ISSN 1990-9241

THEMTAICS is an Effective Tool for Prediction of Structural Genomics

Ihsan A. Shehadi

Department of Chemistry, United Arab Emirates University,
P.O. Box 17751, Al Ain, Abu Dhabi, United Arab Emirates

Abstract: New directions in computational methods for the prediction of protein structures by which ionizable residues with extreme Pka values in proteins and enzymes are explored. Usually, the pKa of such residues fall beyond the normal range (0-14 pH units), however, the reasons for such extreme behavior have not been understood yet. It is strongly believed that residues, act as extreme acids or bases, play a very important in the protein folding process through which turns and separation between loops are controlled. In this research work, THEMATICS, a method for the location and characterization of the active sites of enzymes, is used to locate residues with extreme behavior in Human Adenosine Kinas (KA), Triosephosphate Isomerase (TIM), Phosphate mannose isomerase (1PMI) and Human pepsin (1PSO). THEMATICS, for Theoretical Microscopic Titration Curves, is based on well-established finite-difference Poisson-Boltzmann methods for computing the electric field function of a protein. The chemical basis for the predictive powers of THEMATICS is also featured in this work.

Key words: Electrostatic potentials · computational methods · THEMATICS · extreme pKa's

INTRODUCTION

In the post Genomic era, research holds tremendous promise for untold innovation in medicine, including a host of new kinds of therapeutic agents and diagnostic assays. However, to make this promise a reality, one of the most critical tasks is to discover the structure of the many gene products, particularly enzymes and functional proteins, whose sequences we now know for humans and dozens of other organisms [1-4]. Structural genomics efforts are rapidly increasing and are needed to resolve the number of unknown structures of these gene products. The next important step is to establish an insight into the structure of these proteins in normal and diseased states. Herein, the unique features of the structurally important sites in a set of four proteins are examined how these residues can form the basis for prediction of protein folding through their interaction with other peripheral or active site residues [5-8].

Acid denaturation of protein has been a widely studied phenomenon for many years. It has been shown that some proteins retain substantial native like structure under acidic conditions forming a "molten globules' species [9]. There are great possibilities that these "molten globules" contain highly acidic or basic conserved residues across species and functions. Hence, they (molten globules) may play an important role in protein folding pathways or may provide insights as fundamental free energy relationships in protein folding. A variety of experiments that involved human kinases and myoglobines have indicated the involvement of such residues in salt bridges, hydrogen bonds and disulfide bridges [10, 11]. Due to the vast amount of newly discovered proteins, it is important to seek computational approaches to aid in structural determination of genomic products. This is done by mapping out the residues important in protein folding process and hence explains the relationship between function and structures in proteomics.

However, very few computational approaches are developed to determine the pH and salt dependence of protein stability based on the analysis of three-dimensional structure of proteins in question. THEMATICS (Theoretical Microscopic Titaration Curves Method) approach will be used in this paper to obtain the pKa of all ionizable residues in proteins of interest and finding those ones who possess the property of extreme acids (low pKa) or bases (high pKa).

The main important feature of THEMATICS [12] is locating and characterizing the active sites of proteins using the computed pH-dependent behavior of the ionizable residues in the protein structure. THEMATICS is based on finite difference Poisson-Boltzmann (FDPB) methods for calculating the

Corresponding Author: Dr. Ihsan A. Shehadi, Department of Chemistry, United Arab Emirates University, P.O. Box 17751, Al Ain, Abu Dhabi, United Arab Emirates

electrical potential function for a complex array of charges, coupled with a Monte-Carlo procedure for determining the average charge as a function of pH for the ionizable residues in that protein's calculated potential. These methods have been in use for some time to predict the pK_a's of residues in proteins. FDPB methods solve the Poisson-Boltzmann equations to obtain the electrical potential function for a complex three-dimensional assembly of charges. Such methods only require the three-dimensional structure of the protein as input plus a set of partial charges, obtained from a force field for each atom in the protein structure. The novel feature of THEMATICS is that it extracts information from the shapes of the theoretical titration curves for the ionizable residues in the protein structure, as derived from a FDPB calculation [13, 14].

MATERIALS AND METHODS

THEMATICS; Theoretical Microscopic Titration Curves; is used for locating the active sites of proteins from the three-dimensional structure alone [12-15]. The electrical potential function for the protein is obtained by utilizing the Finite Difference Poisson-Boltzmann methods [24-26], followed by calculating the predicted titration curves for all of the ionizable residues in the protein structure. The shapes of the predicted titration curves are analyzed to identify those residues with non-sigmoid titration behavior. Usually, a cluster of two or more such anomalous residues in physical proximity is a highly reliable predictor of the active site [16-18].

The most important feature of THEMATICS is that it only requires the three-dimensional structure of the query protein as input. The protein in quest does not have to bear any resemblance in sequence or in structure to any previously characterized protein. Curves of the mean net charge as a function of pH (titration curves) are calculated for all of the ionizable residues in each of the model structures. Therefore, pKa's of all ionizable residues will be determined frm the theoretical titration curves.

Then curves are analyzed to select the ones that deviate from the typical sigmoid shape. Most of the curves do possess the characteristic sigmoid shape, with a sharp fall-off in charge in the region around the midpoint, as predicted by the Henderson-Hesselbach equation. Ionisable residues deviate from the typical behavior are the ones identified by THEMATICS to be potentially important. Then searching for clusters of residues with deviant titration behavior that are in a physical proximity.

Arg (R), Asp (D), Glu (E), His (H), Lys (K), Tyr (Y) residues, all Cys (C) residues that are not involved in disulfide bridges, plus the N-and C-termini are considered ionisable. A typical residue to be ionisable in a protein, it must obey the Henderson-Hesselbach equation for the pH as a function of the concentrations of an acid HA and its conjugate base A, given by:

$$pH = pK_a + \log\{[A^-]/[HA]\} \quad (1)$$

The Henderson-Hesselbach can be expressed in terms of the mean net charge C as a function of pH:

$$C(pH) = 10^{pKa}/(10^{pH} + 10^{pKa}) \quad (2)$$

Equation (2) holds for the Arg, His and Lys residues that form a cation upon protonation:

For Asp, Cys, Glu and Tyr, residues that form an anion upon deprotonation. The Henderson-Hesselbach equation in C(pH) form is given by:

$$C(pH) = ?10^{pH}/(10^{pH} + 10^{pKa}) \quad (3)$$

Equations (2) and (3) both have sigmoid shape where C exhibits a sharp decline in the region around the pK_a.

The pKa of an ionisable is predicted when the average charge falls sharply near the midpoint as pH is increased on the titration curve. The ionisable species goes from fully protonated to fully deprotonated in a narrow pH range. On a few occasions in the past it has been noted that a small fraction of the residues in a protein have perturbed titration curves (as computed by a FDPB method) that do not fit this typical pattern [16, 19, 20, 21]. Since the Henderson-Hesselbach equation applies to an uncoupled, monoprotic acid in the absence of a variable electric field, one might expect that equation (1) would break down for certain residues in proteins. It was shown that a particular type of perturbed titration curve, one that has a flat or nearly flat region in the C(pH) function, such that partial protonation persists over a wide pH range, has functional significance [15, 22].

THEMATICS calculations are simple and fast. On a single-processor desktop personal computer, total real time to analyze one protein ranges from less than one hour for the smallest enzymes to about one day for a large multimeric structure with thousands of residues.

RESULTS

The, titration curves for the proposed for four different proteins are calculated using the methods described in detail in reference [12] and are based on a finite-difference Poisson-Boltzmann procedure [16-21]. In all cases illustrated here, we started with the

Table 1: Residues with extreme pKa along with their conservation scores, locations within the protein structure and active site residues of the Human Adenosine Kinase (1bx4)

Residue	pka	Position	Conservation. score out of 9	Active site residue
y112	16.4	Beta sheet -middle	5	NO
R132	16.8	Beta sheet -TURN	8	NO
Y166	16	Beta sheet middle	6	NO
D294	-0.4	TURN	9	NO

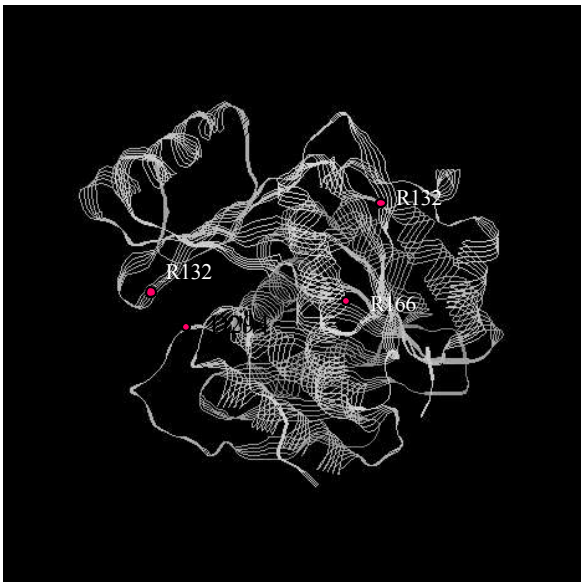

Fig. 1: The secondary structure of the Human Adonisine Kinas (KA) displayed with the approximate locations of the extreme pKa residues

three-dimensional structure of the proteins those are obtained from published x-ray crystallographic data from Protein Data Bank. The Human Adenosine Kinas carries the code of 1bx4, Triosephosphate Isomerase (TIM) has a code of 1TPH, Phosphate mannose isomerase has a code of 1PMI and Human pepsin has the code of 1PSO. The perturbed titration curves that have extended regions of partial protonation for each protein structure may be identified either by simple visual inspection or by a statistical analysis of the curve features. Then tabulated theoretical numerical values of pKa's of the ionisable residues are obtained. At last, the visual analysis of the three dimensional analysis along with the comparative studies using spdbv viewer [23] program are conducted.

In all cases, the calculations are conducted on the native protein structures without any substrates, inhibitors or activators are being bound to the proteins. This is done in order to simulate spontaneous protein folding without any induced conformational changes. The role of the extreme pKa's on the conformational changes of enzymes is beyond the scope of the work and will be studied in future work.

Human adenosine kinas (KA): Human adenosine kinas (KA) catalyzes the transfer of a phosphate group from ATP or GTP to a ribofuranosyl-containing nucleoside analog. AK has a unique fold consisting of an alpha-beta three layered plus a smaller alpha-beta two layers domain. AK is involved in the regulation of adenosine levels (Fig. 1).

The single x-ray structure of human KA is determined from the protein data bank with code 1bx4 and resolution of 1.5 ?A. The active site residues are identified by THEMATICS and from experiment to be D18, D300 and E226. Three extreme base residues (Y112, R132 and Y166) and one extreme acid residue (D294) are identified by THEMATICS. Y112 and Y166 are isolated and located at in the middle of a beta sheet where phenol ring is pointing towards a helix controlling the distance between the helix and the sheet. The locations of both residues are far from the active site ones, hence their direct contribution to activity of the enzyme is not significant. They only play a role in guiding the formation of the loops at the outer distant area of the protein. That explains Y112 and Y166 residues have low conservation scores of 55.56 and 66.67% respectively (Table 1).

On the other hand, residues R132 and D294 have 88.89 and 100 % conservation score respectively. D294 is acting as an extreme acid and is found at one end of a helix where a change in direction of the fold is observed. R132 has a very high calculated ionization constant; this makes it an extreme base. It is located on beta sheet where change in the direction of folding is seen. The average distance between D294 and R132 is 7.55 A. Both are located in vicinity of the active site residues. Therefore, D294 and R132 play a significant role in the activity of the protein by providing the necessary electrostatic potentials and preserving the secondary structure of the active pocket. That could provide an logical explanation why these residues are conserved in the evolutionary tray.

Triosephosphate isomerase (TIM): Triosephosphate Isomerase (TIM) catalyses the conversion of D-glyceraldehyde 3-phosphate (GAP) to Dihydroxyacetone Phosphate (DHAP) and has the α/β barrel ("TIM barrel") fold. The x-ray crystal structure data for TIM from chicken (PDB code 1TPH) is obtained from the Protein Data Bank with a resolution

Table 2: Residues with extreme pKa along with their conservation scores, locations within the protein structure and active site residues of the trio phosphate isomerase (TIM)

Residue	Pka	Position	Conservation score out of 9	Active site residue
D227	-1.2	At the beginning of a sheet	9	NO
E77	-1.8	On aturn	9	NO
E97	-1.8	In a helix	9	NO
E104	-2.2	On aturn	9	NO
K112	16.2	In a middle of a helix	9	NO
R98	18.8	In a helix along with R99	9	NO
R99	17.2	In a helix along with R99	9	NO
R189	16.2	In a middle of a helix	9	NO

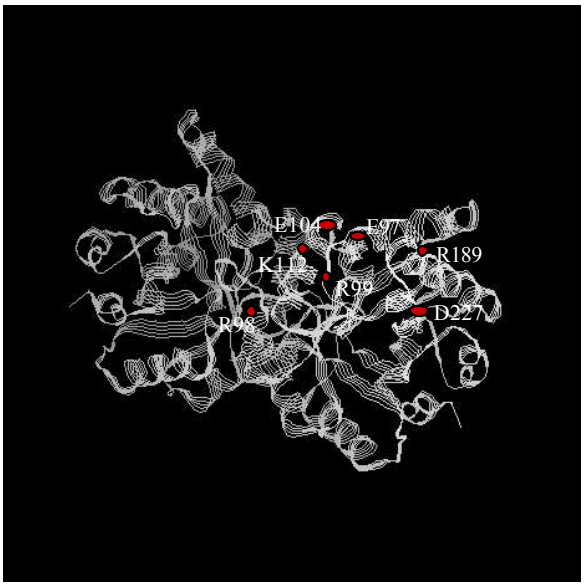

Fig. 2: The secondary structure of the tiophosphate isomerase displayed with the approximate locations of the extreme pKa residues

of 1.8 A. Since TIM is active as a dimmer, the calculations are performed on the dimmer (Fig. 2). THEMATICS identifies H95, E165,C126 and Y164 as the active site residues. However, experimentally only H95 and E15 are identified. Eight other residues are found on each monomer to have negative ionization constants (pKa) and acting as very strong acids (D227, E77, E97 and E104) and other four residues with high ionization constant making them act as extreme bases (R89, R98, R99 and K112). D227 is located at the beginning of a beta sheets before a fold changes direction. D227 (acidic) and R189 (basic) bind together, hence controlling control the distance between the helix and the beta sheets. R189 is located just below a sheet perpendicular to its plane, hence acting as an extra support for the sheet (Table 2)

The two E77 residues, in monomer 1 and 2, have an extreme negative pKa's that make them very acidic. Both residues are highly conserved with a score of 100 % conservation. Structurally, they hold the two monomers together through binding with residues located on different monomers. The distance between OE2-E77 (1) and NZ-K112 (2) and NH1-R98 (1) are measured to be 4.74 A and 2.77 A respectively. The distances between OE2-E77 (2) and NZ-K112 (1) and NH1 (2) are 4.89 and 3.47 respectively. K112 and R98 could easily bind with each other since the distance of separation is calculated to be 5.87 A. E104 is also acting as an acid and within the binding distance with two basic residues K112 and R98. The NZ-K112 and OE1-E104 are 2.92 A? apart and NH2-R98 is 4.38 A from OE2_E104. Therefore, residues E77, R98, E104 and K112 are bound together in a cluster, conserving the shape of folds and the cleft in mid dimmer region through which active site is accessed.

Residues R99 are very basic with a pKa value of 17.2. It is found near the active site residue H95. The distance between NH2-R99 and N-H95 is 5.96A. R99 also residue has 100% conservation according to consurf protocol. E97 is also found near H95 and could bind to it. The distance between EO1-E97 and ND1-H95 is 3.47 A and 3.04A between N-E97 and O-H95. H95 is located in the middle plane between R99 and E97and could bind to residues H95, Y164 and E165, hence controlling their orientations in the active site and preserving the enzyme catalytic activity.

Phosphate mannose isomerase (1PMI): Phosphomannose isomerase (PMI) catalyses the reversible isomerization of fructose-6-phosphate (F6P) and mannose-6-phosphate (M6P). In the absence of PMI activity in yeasts, cell lysis will occur and thus the enzyme is a potential target for inhibition and may be a route to antifungal drugs. The x-ray crystal structure of PMI is determined form the protein data bank with a resolution of 1.7 A and a code of 1PMI. There are three distinct domains in the PMI; the active site lies in the central domain which is flanked by a helical domain on one side and a jelly-roll like domain on the other (Fig. 3).

There are seven residues act as strong bases with extreme high pka's(R8, Y80, K99,R199,R241, R248 and Y287) and five residues act as a strong acids with extreme low pka's (H134, E138, D239, H285 and E48). E48 and K100 can bind to each other through

Table 3: Residues with extreme pKa along with their conservation scores, locations within the protein structure and active site residues of the Phosphate mannose isomerase

Residue	Pka	Position	Conservation. score out of 9	Active site residue
R8	15.5	Beta sheet -middle	8	NO
y80	16.8	Short helix	9	NO
K99	16.4	Sheet-middle	9	NO
H133	-9.1	On a turn	9	NO
E138	-4.4	Beta sheet -middle	9	YES
R199	15.3	Guiding a fold	5	NO
D239	-0.11	Forcing aturn	9	NO
R241	15.1	Short helix II D239	1	NO
R248	15.1	Helis-middle II R241	6	NO
H285	-7.1	Beta sheet -middle	9	YES
E48	0.2	Beta sheet -on a turn	9	NO
Y287	19.1	Beta sheet -II	9	NO

Table 4: Residues with extreme pKa along with their conservation scores, locations within the protein structure and active site residues of the Human Pepsin (1 PSO)

Residue	Pka	Position	Conservation. Score out of 9	Active site residue
Y14	17.01	Beta sheet II lop	9	
Y125	15	On a turn II	6	
R307	20.3	Helix-guiding a turn	9	
R315	16.08	Beta sheet -guiding a turn	8	

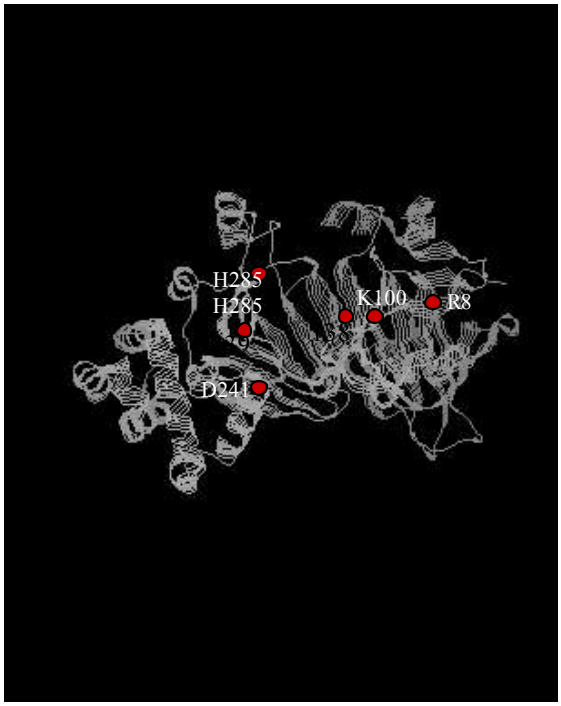

Fig. 3: The secondary structure of the phosphate mannose isomerase displayed with the approximate locations of the extreme pKa residue

NZ-R48 and OE1-K100 where the distance measured to be 2.89 A. K100 is located in the middle of a beta sheet where it is very close to E48, hence forcing an opposing beta sheet (where the E48 is located) to turn. E138 and H285 are both active site residues with extreme low pka's values. The orientation of E138 is conserved since it binds to Y287 and H285. The distance between OE1-E138 and OH-Y287 is 4.62 and between OE2-E138 and NE2-H285 is 3.90 A. Therefore the catalytic activity of residue of E138 is maintained through preserving it s orientation in the pocket of the active site. R8 residue is located in the middle of a beta sheet point towards the beginning of a helix. The distance between NH2-R8 and O-S34 is 2.67 A which makes the two residues bind to each other. R8 play a significant role in guiding the direction by which the helix will be forming. Y80 is another basic residue that is highly conserved among species. It is found to be on one turn helix and extends parallel to the terminus part of the protein. It binds to E4 since the distance between OH-Y80 and OE2-E4 is 2.43 A. H135 is an extreme acidic residue with pka of-9.1. It forms a salt bridge with E357 through binding between OE1-E357with NE2-H135 and OE2-E357 with NE1-H135 that are separated by 4.78 and 3.67 A respectively. It is essential in controlling the distance between two sheets near the active site pocket, hence contributes to preserving the fold around the active site motif. Moreover, H135 and E138 are found in the same beta sheet but pointing in opposite directions forcing E138 to be exposed inside the pocket. R199 is partially conserved residue across species with a score of 50 %. It is important in shaping the fold away from the active site, does not affect the catalytic activity of the protein. R199 is forming a salt bridge with D195 and could bind to TH322 since the distances between OD2-D195:NE-R199, OD1-D195:NH2-R199 and O-T322:NH1-R199 are 3.15, 3.47 and 2.85 respectively. D239 is located on a loop where

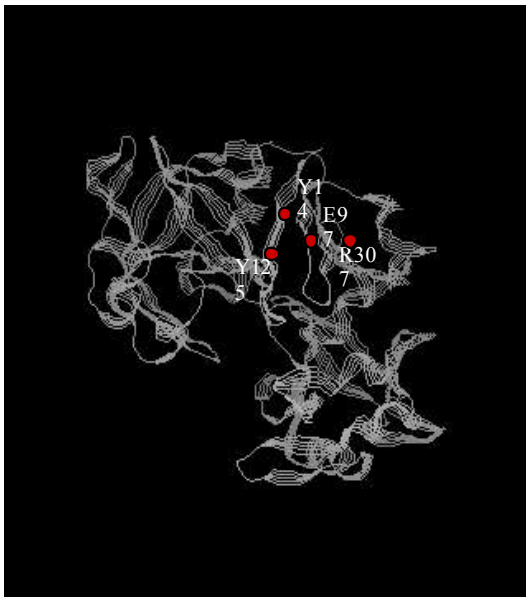

Fig. 4: The secondary structure of the human pepsin displayed with the approximate locations of the extreme pKa residue

a turn is forced. The conservation score for D214 is 11.11 % only. Therefore, this residue plays a role in folding the peripheral part of the protein far away from the active site pocket. It forms two hydrogen bonds with H268 which is also identified as a false positive by THEMATICS. Therefore, the distance between a loop and a helix is maintained. R248 is also has high pka and is located exactly in the middle of helix and pointing in the same direction as the helix. Therefore, it plays a role in controlling the length of a helix (Table 3).

Human pepsin (1PSO): The three-dimensional crystal structure of 1PSO (pepsin) had been solved by X-ray crystallographic methods. The native pepsin structure has been refined with data collected to 2.2 A resolution to an R-factor of 19.7%. The enzyme undergoes a conformational change upon inhibitor binding to enclose the inhibitor more tightly. The analysis of the binding sites indicates that they form an extended tube without distinct binding pockets. (Fig. 4)

There are four basic amino acids residues Y14, Y125, R307 and R315 with high pKa's are located in the vicinity of the active site pocket. R307 is the only active site residue with high pka due to its direct role in the chemical activity of the protein. Y125 is located on a turn of a loop while R315 is found to be on helix where a change in direction is observed. Both residues are very basic with very high pka's and highly conserved across species with a conservation score of 100% of the seventeen species studied. The OH-Y125:N-R315 distance is 5.68 A and O-Y125:NH1-R315 distance is 6.20 where hydrogen bonds can be formed. Both residues are also located at the entrance to the active site pocket and facing towards each other, hence the size of the cavity is controlled through the previously mentioned hydrogen bonds formation. R307 and Y14 are highly conserved and have relatively high conservation through species is crucial to preserve the protein function (Table 4)

R307 and Y14 are also located at the other end of the active site. They play the same role as R315 and Y125 in preserving the protein function.

CONCLUSIONS

It is shown that the extreme pKa's residues play very important roles in controlling the secondary structure of proteins either by guiding a turn of a loop and by controlling the distances between different loops, sheets and helices within proteins. In addition to structural integrity, extreme pKa residues also contribute to preserving the catalytic activity of the active site by controlling the shape of the catalytic motif, orientation of active site residues and conserving the access routs by which substrates reach the active site pockets. It is shown that residues with exterem pKa's are involved in salt bridges and hydrogen bonding. Finally, all extreme pKa's residues are highly conserved, which proves their importance in preserving structure as well as functions of enzymes.

ACKNOWLEDGEMENT

This work was financially supported by the Research Affairs at the UAE University under a contract no. 05-03-2-11/05.

REFERENCES

1. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, 2000. Nucleic Acids Res., 28: 235-242.
2. Westbrook, J. *et al*., 2003. Nucleic Acids Res., 31: 489-491.
3. Pearson, W.R. and D.J. Lipman, PNAS, 85: 2444-2448.
4. Pearson, W.R., 1990. Methods in Enzymology, 183: 63-98.
5. Thompson, J.D., D.G. Higgins and T.J. Gibson, 1994. Nucleic Acids Res., 22: 4673-4680.
6. Fiser, A., R.K. Do and A. Sali, 2000. Protein Sci., 9: 1753-1773.
7. Schwede, T. *et al*., 2003. Nucleic Acids Res., 31: 3381-3385.

8. Yang, A. and B. Honig, 1994. J. Mol. Biol., 237: 602-614.
9. Hayward, H., 2004. J. Mol. Biol., 329: 602-614.
10. Tan, T., M. Oliveberg, B. Davis and R. Fersht, 1995. J. Mol. Biol., 254: 980-992.
11. Krupa, A. G. Preethi and N. Srinivasan, 2004. J. Mol. Biol., 329: 1025-1039.
12. Ondrechen, M.J., J.G. Clifton and D. Ringe, 2001. Proc. Natl. Acad. Sci., USA, 98: 12473-12478.
13. Shehadi, I.A., A. Uzun, L. Murga, V. Ilyin and M. J. Ondrechen, 2005. J. Bioinformatics and Computational Biol., 3: 127-143.
14. Shehadi, I.A., H. Yang and M.J. Ondrechen, 2002. Mol. Biol. Reports, 29: 329-335.
15. Ondrechen, M.J., J.M. Briggs and J.A. McCammon, 2001. J. Am. Chem. Soc., 123: 2830-2834.
16. Bashford, D. and K. Gerwert, 1992. J. Mol. Biol., 224: 473-486.
17. Bashford, D. and M. Karplus, 1993. Proteins, 15: 266-282.
19. Sampogna, R.V. and B. Honig, 1994. Biophys. J., 66: 1341-1352.
20. Carlson, H.A., J.M. Briggs and J.A. McCammon, 1999. J. Med. Chem., 42: 109-117.
21. Beroza, P., D.R. Fredkin, M.Y. Okamura, G. Feher, 1995. Biophys. J., 68: 2233-2250.
22. Shehadi, I., 2003. 3rd Pacific Symposium on Biocomputing. Hawaii, USA.
23. http://www.expasy.org/spdbv/