ELSEVIER

Biophysical Chemistry 105 (2003) 595–608

Biophysical Chemistry

www.elsevier.com/locate/bpc

Protein structural alignment for detection of maximally conserved regions

Vladimir Kotlovyi[a], William Lee Nichols[b], Lynn F. Ten Eyck[a,*]

[a]*San Diego Supercomputer Center, University of California, San Diego 0505, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA*
[b]*Department of Chemistry and Biochemistry, University of California, San Diego 0645, 9500 Gilman Drive, La Jolla, CA 92093-0645, USA*

Abstract

An algorithm for comparison of homologous protein structures and for study of conformational changes in proteins, has been developed. The method is based on identification of pieces of the two molecules that have similar shapes, as determined by the local conformation of the polypeptide chain. Pieces that superpose within a specified tolerance are assembled into domains based on similar transformations for superposition. The result is sets of pieces that represent conserved structural elements and conserved spatial relationships between structural elements within the proteins being compared. A similarity criterion based on maximum distance rather than on root mean square deviation reduces bias by outliers. The utility of the method is demonstrated by using examples from the protein kinase family.

Keywords: Protein structure comparison; Protein structure alignment

1. Introduction

In the course of our research on the structural changes in hemoglobin on oxygenation [1] and on the structure and function of members of the protein kinase family of enzymes [2–9] we have developed an approach to structural alignment of proteins optimized for finding areas of highly conserved structure. In the case of hemoglobin [10,1] we noted that the structural changes in the transition between the oxygenated and deoxy forms of the molecule preserve a 'rigid core' that consists of half of the molecule. This core is nearly twice as large as initially suggested and provides a possible mechanism for communication of stereochemical changes across large distances. Analysis of such conformational change is greatly simplified if the portions of the structure that are invariant are easily identified.

The protein kinase family represents a very large and diverse family of closely related proteins. Sequence analysis demonstrated many conserved motifs, some common to all protein kinases and others as specific signatures of particular subfamilies. Now that there are more than 150 protein kinase structures in the Protein Data Bank [11] it is clear that various motions and rearrangements of these motifs are essential for activation, regulation and specificity of these vital enzymes. It is also clear that there are many mechanisms for regulating these processes.

*Corresponding author. Tel: +1-858-534-5141; fax: +1-858-822-0873.
E-mail address: lteneyck@ucsd.edu (L.F. Ten Eyck).

0301-4622/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0301-4622(03)00069-3

Performing the same kind of structural conservation analysis on families of proteins as can be done for allosteric proteins requires a structural alignment to determine sets of 'equivalent' residues. Programs and databases developed for fold recognition and homology detection, such as CATH [12], CE [13], DALI [14], ProSupp [15], SCOP [16,17], PrISM [18] and VAST [19,20] provide useful information but are not optimal for this purpose. MUSTA [21] is directed towards the problem of multiple structural alignment, and uses a notion of transformational clustering similar to that described in this work. The primary distinction between these resources and our work is that these methods are tuned for finding as much commonality as possible, where we are interested in finding the smaller amount of very tightly conserved structural elements. Furthermore, these methods return only the 'best' alignment, but 'best' does not have a unique definition in this context.

Our approach to analysis of protein structure assumes that protein structures often have very well determined portions that can be considered structurally invariant, at least over the range of conditions of interest to us. We also note that many protein structures contain portions that are more malleable, whether this is due to inherent flexibility of loops and surface-exposed amino acid residues, or due to reorientation of rigid bodies by hinge motions. The specific chemical and biological properties of particular proteins depend on both rigidity and flexibility.

To examine these properties further we have developed a structural alignment program that searches first for fragments with matching shape. It then assembles these fragments into alignments subject to a maximum allowed distance between matched pairs of atoms. The algorithm is called SAT (Shape And Transformation). Mathematically this problem does not have a unique answer. The program returns all of the alternative alignments for evaluation by the user. The multiple solutions provide valuable information about the nature of the conserved structures.

2. Methods

The goal of this work is to find conserved shapes in proteins. We therefore start with a direct measure of shape rather than a simple measure of distance (RMSD, root mean square deviation) between two structures. Taking the sequential trace through the C_α atoms as a curve in space we look for matching curves. The key issue is to compare shape by using metrics that are invariant with respect to orientation and position, and which are completely descriptive of the shape. We have developed two procedures, one based on differential (infinitesimal) invariants and the other based on finite analogs of these infinitesimal invariants derived from the virtual bonds between successive C_α atoms.

2.1. Differential geometric approach

The premise of the differential geometry approach is that two proteins or structures therein exhibit primary structure alignment where smooth space curves fit through C_α atom coordinates of each protein are identical. Depending only upon structurally representative space curves, this approach to structure-based alignment is thoroughly blind to primary sequence and side-chain orientation and chemistry.

The differential geometric representation of a space curve given by arc length, curvature and torsion $(s(t),\ \kappa(s)(t)),\ \tau(s(t)))$ and its Cartesian parametric representation $(x(t),\ y(t),\ z(t))$ are equivalent. A fundamental result of the differential geometry of space curves is that two space curves are congruent if their differential geometric coordinates are the same [22]. Structurally aligning two different protein main-chain conformations entails finding space curve segments in each with similar geometry.

The method is illustrated by plotting the curvature $\kappa(s)$ and torsion $\tau(s)$ of a protein space curve representation as two ordinates with the same abscissa s, the arc length. The first and third graphs in Fig. 1 show $\kappa(s)$ and $\tau(s)$ for Protein Kinase A (PKA). The differential geometric representation for one protein conformation can be overlaid upon that of another to find intervals in arc length with similar curvature and torsion. The shifts in arc length necessary for matching intervals of one differential geometric representation with another define insertions and deletions needed to align the

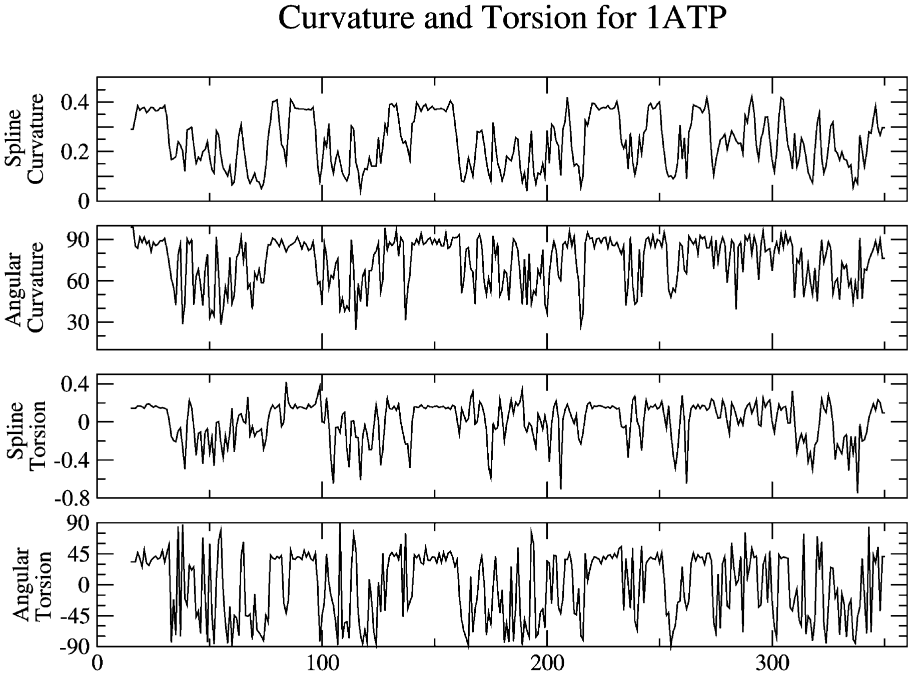


Fig. 1. The top graph and the third graph show the curvature and torsion for the interpolating spline curve passing through the C_α coordinates of the catalytic subunit of Protein Kinase A (PKA). The second and fourth graphs show the corresponding quantities in terms of C_α–C_α virtual bond angles and torsion angles, in degrees. Although the angular plots appear noisier, the information content is essentially identical. Spearman's rank correlation coefficient between the angular values and the spline values is 0.97 for curvature and 0.95 for torsion.

two conformations with each other. The curvatures of two protein space curve representations are defined as similar over intervals of arc length if the maximum absolute value of the difference in curvature does not exceed a prescribed value over the interval of arc length. Similarity of torsion is analogously defined.

Calculation of curvature and torsion is based on the use of cubic spline functions to describe a smooth curve through the C_α coordinates ([23], Section 3.3). This is the same function used in many molecular graphic programs. The spline function is conveniently described in parametric form as $\mathbf{r}(x(t), y(t), z(t))$ where $x(t)$, $y(t)$ and $z(t)$ are piecewise cubic polynomials of t. As t sweeps through its range of values the function $\mathbf{r}(x(t), y(t), z(t))$ passes through the C_α coordinates in order.

The arc length $s(t)$ is defined as

$$s(t) = \int_0^t \left| \frac{d}{du}\mathbf{r}(x(u), y(u), z(u)) \right| du$$

The space curve now can be parametrized with s instead of t. The unit tangent vector $\mathbf{u}(s)$ is given by

$$\mathbf{u}(s) = \frac{d\mathbf{r}(s)}{ds}$$

The curvature $\kappa(s)$ is

$$\kappa(s) = \left| \frac{d\mathbf{u}(s)}{ds} \right| = \left| \frac{d^2\mathbf{r}(s)}{ds^2} \right|$$

Finding the torsion $\tau(s)$ of the space curve representing the protein conformation requires two more vectors, the unit principal normal vector $\mathbf{p}(s)$

and the unit binormal vector $\mathbf{b}(s)$.

$$\mathbf{p}(s)=\frac{1}{\kappa(s)}\frac{d\mathbf{u}(s)}{ds}$$

$$\mathbf{b}(s)=\mathbf{u}(s)\times\mathbf{p}(s)$$

The torsion $\tau(s)$ is then

$$\tau(s)=\mathbf{p}(s)\cdot\mathbf{u}(s)\times\frac{d\mathbf{p}(s)}{ds}=-\mathbf{p}(s)\cdot\frac{d\mathbf{b}(s)}{ds}$$

The functions $\kappa(s)$ and $\tau(s)$ are independent of orientation and translation, and completely characterize the shape of the curve described by $\mathbf{r}(x(t), y(t), z(t))$.

2.2. Angular invariants

An alternative representation of the shape of the polypeptide chain is derived from a representation in internal coordinates, where the molecular structure is expressed in terms of bond lengths, bond angles and bond torsion angles. The application of this description to protein structure in terms of C_α–C_α virtual bonds was pioneered by Rackovsky and Scheraga [24]. The length of the C_α–C_α virtual bond is essentially constant in protein structures, so the only relevant variables are the angles between successive C_α virtual bonds and the torsion angles about the C_α virtual bonds. The torsion angles are defined as the angle between the normals to the planes defined by the atoms ($C_{\alpha,i-2}$, $C_{\alpha,i-1}$, $C_{\alpha,i}$) and the atoms ($C_{\alpha,i-1}$, $C_{\alpha,i}$, $C_{\alpha,i+1}$). Clearly the torsion angle is undefined for $i<3$ or $i>N-2$, where N is the number of amino acid residues in the chain, and the virtual bond angles are not defined for $i=1$ or $i=N$. The bond angles and torsion angles are finite analogues of the differential invariants described above.

2.3. Identification of structurally conserved fragments

Regardless of the method used to define the secondary structure signature of the protein chain, the next step in the process is to locate sections of the two chains being compared that have similar secondary structure. This is done by starting with a short fragment of consecutive residues in the ‘probe’ chain, which is the chain to be mapped onto the ‘target’ chain. The secondary structure signature (i.e. the values of torsion and curvature or bond angle and torsion angle) of the probe fragment is moved down the target chain and compared at each step with the target signature. A match is defined as absolute differences in torsion and curvature below a predefined threshold. The algorithm proceeds as follows.

1. Compare the probe signature with the target signature. If a match is detected do steps a and b.
    a. Extend the match by looking ahead in both the probe and target secondary structure signatures. Find the longest fragment that still matches.
    b. Compare the boundaries of the fragment with the boundaries of fragments previously found. If the new fragment is completely contained in a previously found fragment, discard it; otherwise add it to a list of possible components.
2. Step the probe one residue down the target and repeat step (1) until the end of the target chain is reached.
3. Adjust the probe fragment by moving one step down the probe sequence. If the end of the probe chain has not been reached, restart step (1) at the beginning of the target.

This process is exhaustive. It will find all possible matches, including repeated structural motifs or domains.

2.4. Domain assembly

The result of the previous step is a list of pairs of fragments. Each member of the pair has the same secondary structure signature within the specified tolerance. Each pair is also characterized by a coordinate transformation that will place the probe fragment onto the target fragment. Pairs which have the same coordinate transformation are candidates for combination into a domain. The combining process is as follows.

1. For each pair *i*,
   a. Compute the centroids $\bar{\mathbf{x}}_{i,p}$ and $\bar{\mathbf{x}}_{i,t}$ of the probe and target fragments respectively, and the 3×3 rotation matrix $\mathbf{R}_i$ which rotates the probe onto the target when the centroids are superposed.
   b. Compute the distances from the centroid to each atom for all atoms in each fragment, and save the largest as $\mathrm{d}_{i,p,m}$ and $d_{i,t,m}$.
2. For all pairs of fragments,
   a. Compare the distances between centroids of the target fragments and the centroids of the probe fragments. Reject the pair if these distances are not within MAXD/2.
   b. Reject the pair if $|d_{i,t,m} - d_{j,p,m}| > \mathrm{MAXD}/2$.
   c. Compute the angular distance between the two transformations using the well-known formula for the angle of rotation produced by application of a rotation matrix [25].

$$D_{i,j} = \arccos\left(\frac{\mathrm{trace}(\mathbf{R}_i^{-1}\mathbf{R}_j) - 1}{2}\right)$$

Since **R** is a rotation matrix, the inverse is simply the transpose. If max $(d_{i,p,m},\ d_{i,t,m},\ d_{j,p,m},\ d_{j,t,m}) \times 2\ \sin\ (D_{i,j}/2) > \mathrm{MAXD}$, reject the pair. In other words, reject the pair if the distance between the rotated positions will be greater than MAXD.

The surviving pairs are all potential members of domains; the problem is to assign them to the appropriate domain. Note that a given pair may be a member of more than one domain. The algorithm we use for assigning pairs to equivalence classes is close to that given in section 8.6 of *Numerical Recipes in C* [23], and also described in *The Art of Computer Programming* [26].

Each equivalence class is superposed as a whole using the method of Kabsch [27,28]. This gives

Table 1
Tight structural alignments of monellin and cystatin. The structural alignment generated by CE for monellin and hen egg white cystatin extends from 1MOL:A(8-90)and from 1CEW:I(16-116). Allowing for gaps it aligns 74 out of 94 residues with an *RMSD* of 2.1 A. The largest tight alignment by the SAT method is shown in **BOLD CAPITALS .**

```
       10         20        30          40        50        60
        hh hhhhhhhhhhhhhhh         sssssssssssssss   sssssssssssss
1MOL  igpf-tqnlgkfAVDEENKIGqyg--rltFNKVIrpcMKKTIyenereikGYEYQLYVYA
1CEW  endeglqralqfAMAEYNRA-SndkyssrVVRVIs-aKRQLV-------sGIKYILQVEI
        hhhhhhhhhhh hhhh       ssssssssss sssss          sssssssss
         20        30         40         50               60

                                  70        80        90
                                ssssssssss     sssssss
1MOL  S-----------------------D-KLFRADISEDyktrgrkLLRFNGP
1CEW  GrttcpkssgdlqscefhdepemakYTTCTFVVYSIpwlnqikLLESKCQ
      ssssss     hhhhhhhh      sssssssssss    sssssssss
        70        80        90       100       110
```

Table 2
Geometric comparison of CE and SAT aligned segments

1CEW:I/1MOL:A	Length	CE		SAT	
		RMSD	$d_{\max}$	RMSD	$d_{\max}$
28–36/19–27	9	1.48	2.15	1.47	1.97
44–48/34–38	5	2.39	2.76	1.36	1.88
51–55/42–46	5	1.75	2.16	1.22	1.79
57–67/55–65	11	1.18	1.88	0.66	1.18
92–102/66–76	11	1.12	2.69	1.00	1.43
110–116/84–90	7	1.59	1.95	1.49	1.93

The CE alignment covers a much larger portion of the molecule, while the SAT alignment highlights the regions of greater similarity. The table gives the RMSD and $d_{\max}$ in Ångstrøms between corresponding C_α atoms for each of the segments in bold type shown in table 1.

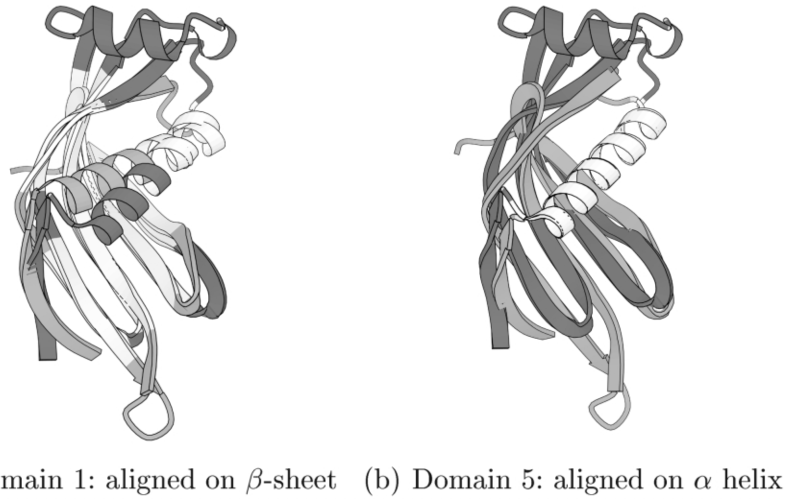
all possible alignments within the limits of the parameter MAXD, and within the limits imposed by the shape similarity criteria. Since there will be no atoms within the set that deviate by more than MAXD, the RMSD is necessarily less than or equal to MAXD.

3. Results

Two systems are presented to show the distinction between this method of structural alignment and others. The first system examines a fold similarity first reported by Murzin [29] between monellin and cystatin. The second system consists of two protein kinases, one of them in two conformations.

Monellin is an extremely sweet protein from the African serendipity berry. The crystal structure used here is PDB entry 1MOL [30]. Murzin [29] noted that the fold is very similar to the cysteine protease inhibitor cystatin (1CEW [31]). SAT finds several alignments shown in Tables 1 and 2, and as MOLSCRIPT [32] diagrams in Fig. 2. This rather obvious example shows that four of the five strands of the β-sheet are almost identical in both structures. It is also possible to superpose the α-helix essentially perfectly, but at the expense of the β-sheet. The RMSD of the two superpositions are 1.19 Å for the superposition of β-structure and 0.51 Å for the superposition of the α-helix.

A more detailed example is given by the comparison of two kinases, the cyclic AMP-dependent protein kinase (PKA), and CK2, formerly known as casein kinase II. The structures used in this study are 1ATP [33] and 1F0Q [34], both closed form kinases. Fig. 1 shows the curvature and torsion for 1ATP, calculated using the spline fit,



(a) Domain 1: aligned on β-sheet (b) Domain 5: aligned on α helix

Fig. 2. Structural alignments of 1MOL (monellin, a sweet tasting protein) onto 1CEW (cystatin, a cysteine protease inhibitor). Monellin is shown in light grey, while cystatin is shown in dark grey. Regions which superpose are shown in white. The β-sheet superposition in (a) has an RMSD of 1.2 Å and covers 48 out of 94 residues. Superposition of the α-helix (b) loses the β-sheet superposition entirely at a MAXD setting of 2.5Å.

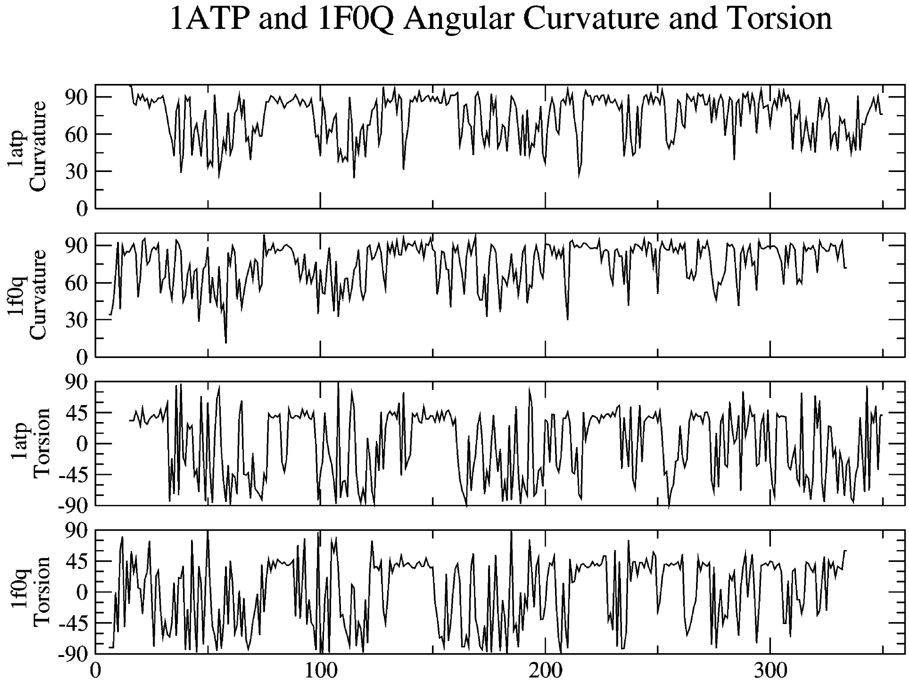


Fig. 3. Comparison of secondary structure signatures of PKA and CK2 shows strong similarities out to the region of residue 250. The signatures of α-helices (curvature near 90° and torsion near 45°) are particularly obvious.

and the virtual bond angles and torsion angles. The curvature graph is similar to the virtual bond angle graph, but there are some differences observable in the comparison between torsion of the spline curve and the torsion angles of the virtual bonds. This difference is superficial. Spearman's rank correlation coefficient [35,36] defined as

$$r_s = \frac{\sum_{i=1}^{n} (\mathrm{rank}(x_i) - (n+1)/2)(\mathrm{rank}(y_i) - (n+1)/2)}{n(n-1)(n+1)/12}$$

measures the correlation between the ranks of two series of numbers. In this case we have $r_s = 0.97$ for the rank correlation between spline and angular measures of curvature, and $r_s = 0.95$ for the rank correlation between spline and angular measures of torsion. The two measures show nearly perfect correlation, and hence contain the same information.

Fig. 3 shows the comparison of secondary structure signatures for 1ATP and 1F0Q. The striking similarity is evident, as are the shifts introduced by insertions and deletions in the alignment. The sequence alignment of 1ATP with 1F0Q, obtained from the Protein Kinase Resource [8], is shown in Table 3. The sequence alignment matches the structural alignment very well in this highly conserved family (Fig. 4).

The next example demonstrates the use of SAT to analyze conformational changes. Protein Data Bank entry 1CMK contains the structure of PKA in an open conformation, with no ATP bound [6]. Superposition of this structure onto 1ATP produces two families of structural alignments, shown in Fig. 5. One family covers primarily the large lobe of the kinase catalytic core; in this case the superposition covers 260 out of 336 residues with an RMSD of 0.86 Å. The second family contains primarily small lobe residues, covers 175 out of 336 residues, and has an RMSD of 1.06 Å. There are a number of residues in common to the two domains, which indicates a possible hinge motion as a close approximation to the relative motion of

Table 3: Sequence and structural alignment of Protein Kinase A (PKA, PDBID 1ATP) and CK2 (PDB ID 1F0Q). The sections marked in CAPITALS are only aligned in domain 1, while those marked in *ITALIC* are only aligned in domain 2. Sections present in both alignments are shown in **BOLD**. Thetwo sequences are 21% identical through the protein kinase catalytic core, which extends from PKAresidue 43 to 297. The optimal sequence alignment didnot correctly locate the large insertion in CK2.

PKA: 1	gnaaaakkgseqesvkeflakakedflkkwetpsqntAQLDQFDRI**KTLG**
CK2: 6	mskarvyadvnvlrpkeywdy--ealtvqw-------GEQDDYEVV**RKVG**
PKA: 51	**T**gs**FGRVM**LVKHkesGNHYAMKILdkqkvvklkqiehtl**NEKRIL**QAV-n
CK2: 47	**R**gk**YSEVF**EGINvnnNEKCIIKIL---kpvkkkkik**R---EIKIL**QNLcg
PKA:100	FP**FLVKLEFS**fkdnsNL**Y--MVME**YVAggemfshlrrigrf_SEP_**HARFYA**
CK2: 91	GP**NIVKLLDI**vrdqhskTP**SLIFE**YVNntd-fkvly--ptl_TDY_**DIRYYI**
PKA:148	**AQ**_IVLT_**FEYLHSL**_D_**LIYRDLKPENLL**dqq-G**YIQVTDFGFAKRV**K-grt
CK2:138	**YE**_LLKA_**LDYCHSQ**_G_**IMHRDVKPHNVM**IdhelR**KLRLIDWGLAEFY**Hpgke
PKA:196	wtL**CGTPEYL-APEIILS**-_K_g_YNKA_**VDWWALGVLIY**EMAagyppff----
CK2:188	ynvR**VASRYFKGPELLVD**_L_qd_YDYS_**LDMWSLGCMFA**GMIfrkepffyghd
PKA:240	-adqpiqiyeki-vsgkvrfpsh_FSSDLKDLLRNLLQV_dltkrfgnlkng
CK2:238	nhdqlvkiakvlgtdglnvylnkyrieldpqlealvgrhsrkpwlkfmna
PKA:288	vnd-ikn_HKWFAT_tdwiaiyqrkveapfipkfkgpgdtsnfddyeeeeir
CK2:288	dnqhl_VSPEAIDFLDKLLRY_dhqerltaleamt_HP_----_YF_-----_QQ_vr
PKA:337	vsinekcgkeftef
CK2:329	aaensr-----tra

the two domains. The residues in common are close to the axis of the hinge, and thus the relative displacement of these residues as a result of the conformational change is small.

The final example compares CK2 with the open form of PKA. This example has structural differences due to both the insertions, deletions and differences in residues shown by the sequence alignment in Table 3 and the difference between open and closed kinase conformations. The results in Fig. 6 show that the SAT alignments can find strongly conserved regions of structural similarity despite both these factors. The largest ‘large lobe’ domain covers 132 residues to an RMSD of 1.19 Å, while the ‘small lobe’ domain cleanly picks out the conservation of the β-sheet in the small lobe. The alignment based on sequence, using a globally optimal method [37], failed to properly align the portion of CK2 following the insertion. The correct alignment was found by the structural alignment procedure.

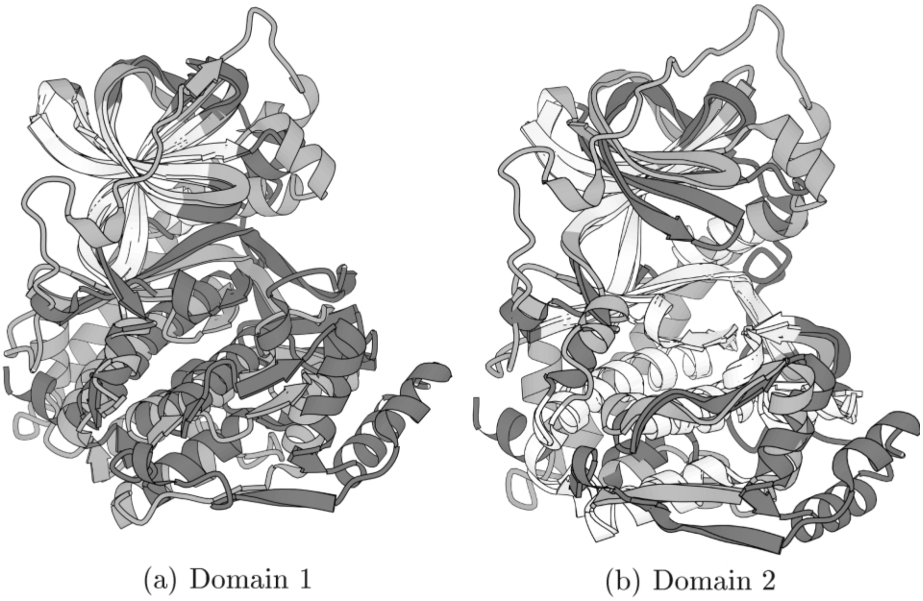


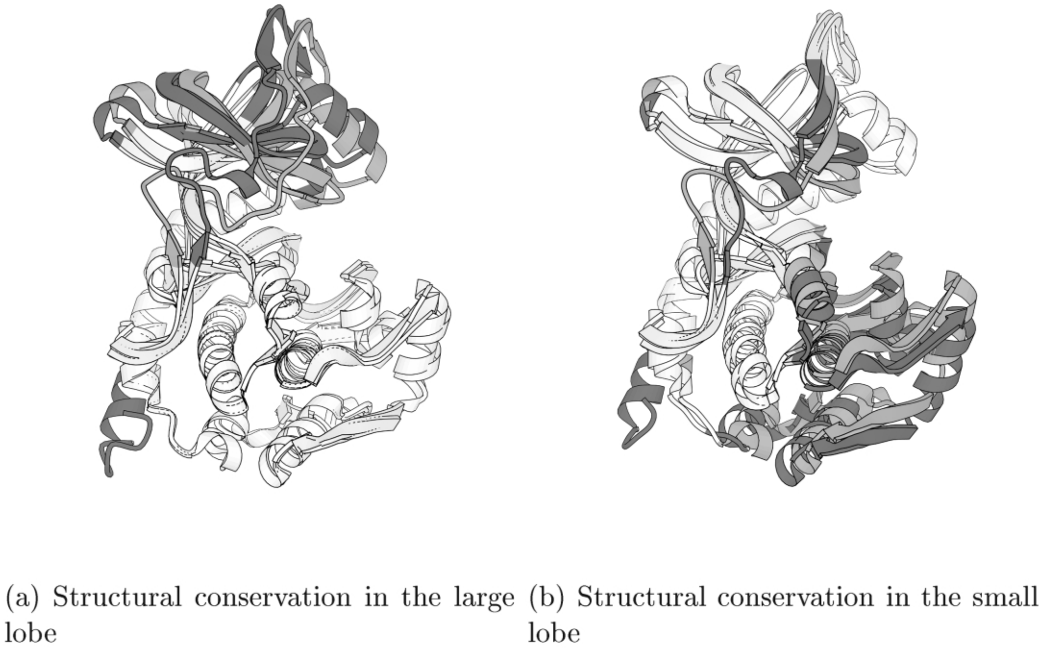
Fig. 4. Structural alignments of 1ATP (PKA) and 1F0Q (CK2). PKA is shown in light grey and white, while CK2 is shown in dark grey and white. The aligned regions are white. The RMSD for Domain 1 is 1.15 Å while that for Domain 2 is 1.08 Å. The close alignment of the β sheet at the top of the structure on the left, the alignment of the helices at the bottom of the structure on the right, and the common portions in the middle of the molecule (aligned in both domains) all show that the two domains are related by a small rotation.

4. Discussion

The algorithm presented here is tuned for the detection of high similarity both locally and in overall packing of a protein structure. This property of the SAT alignments is demonstrated by comparison with a structural alignment given by the CE algorithm [13], which is tuned for the detection of global similarities. Table 1 shows the CE alignment for monellin and cystatin. This alignment covers 74 residues out of 94 with an RMSD of 2.1 Å. SAT breaks the alignment into two major components, one of which covers the β-sheet structure and a small part of the major helix, and the other of which covers all of the α-helix but none of the β-structure. The residues aligned by SAT against the β-structure are highlighted in Table 1. The SAT alignment is not quite a subset of the CB alignment; there are two differences at the ends of SAT-aligned fragments. Specifically, SAT aligns Ser 1CEW:I36 with Gly 1MOL:A27, where CE aligns it with Gln 1MOL:A28, and SAT aligns Asp 1MOL:A66 with Tyr 1CEW:I92 but CE aligns it with Lys 1CEW:I91. These differences both reflect the fact that SAT is considering shape as well as RMSD in computing its alignments.

Table 2 shows that the SAT alignments do in fact pick out highly conserved structural arrangements. The table shows the RMSD and the maximum distance between aligned C_α positions using both the CE alignment and the SAT alignment. In all cases the SAT alignments have lower values by both criteria. Fig. 1a clearly shows how different the orientation of the α-helix is when the structures are aligned on the conserved β-structure. The CE alignment includes the entire helix (Table 1) as well as the β-structure. This shows the similarity of the overall fold. In essence CE captures the architecture of the molecule, while SAT captures the engineering details.

The development of the SAT methodology came from earlier analysis of allosteric proteins [10,11]. A primary goal of this work is to develop tools



(a) Structural conservation in the large lobe

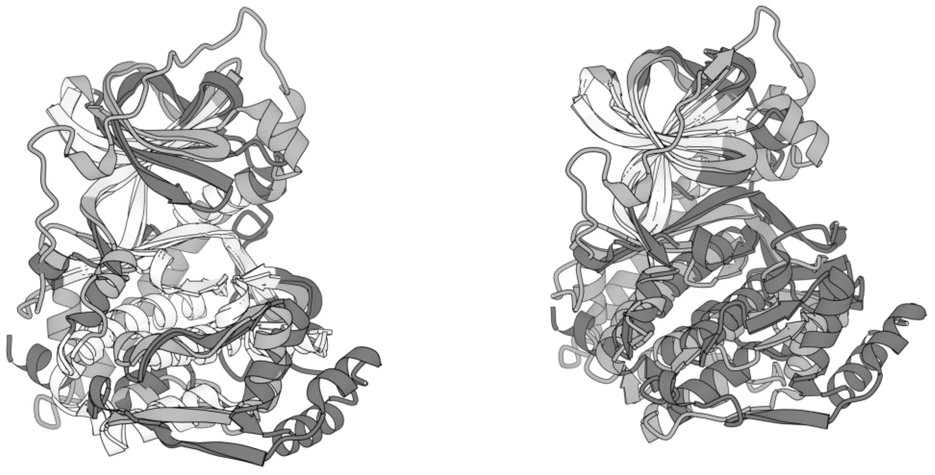
(b) Structural conservation in the small lobe

Fig. 5. SAT finds two sets of superpositions for the open form of PKA (dark grey and white) aligned with the closed form (light grey and white). The domain in (a) aligns 260 out of 336 residues with an RMSD of 0.86 Å. The domain in (b) aligns 175 out of 336 residues with an RMSD of 1.06 Å. The region in the lower left of each figure, common to both alignments, is close to the hinge axis about which the structure opens.

for analysis of structural variability within structural families and during normal function within one molecule. The protein kinase examples demonstrate how well this works. The protein kinase catalytic core has a remarkably well-conserved structure and a very distinctive sequence motif. Protein kinases are usually part of the cellular regulatory process, and thus have to be switched on and off under rigorously controlled circumstances. Activation and substrate binding of these fascinating molecules involves both chemical and conformational changes.

PKA was recognized as a template for the structures of other protein kinases as soon as the structure was determined [2,3]. The structure has been described in terms of 12 subdomains [38]; an introductory description with graphics can be found at the Protein Kinase Resource http://kinases.sdsc.edu as the ‘structure walkthrough’ of PKA. Structural alignment of PKA in the closed form (PDB identifier 1ATP) and CK2 gives two large domains and a set of small fragments. The two largest domains are shown in Table 3. Together these two domains pick out significant elements from all of the subdomains described in the ‘walk-through’. Notably conserved are the glycine loop (43–64), the C helix (89–96), the framework of the ATP binding site (114–124), the entire bottom of the active site and supporting structures (139–192, including the catalytic loop from 164 to 169), and two helices (199–232). CK2 has a 38 residue insertion relative to PKA at PKA residue 255, but structural alignment is recovered for two more helices. Examination of Fig. 4 and Table 4 shows that the two domains are very closely related, suggesting that it might be useful to consider the union of the two domains as a more comprehensive alignment. However, this operation would mask the fact that there is a small but consistent shift of the relative positions of the β-sheet and the lower lobe between the two structures.

The comparison of 1ATP (PKA closed form) and 1CMK (PKA open form) clearly demonstrates the utility of these tools in separating motion from



(a) Structural conservation in the large lobe (b) Structural conservation in the small lobe

Fig. 6. SAT can find significant domains in related proteins despite substantial structural variability compounded by conformational change. The alignment of the open form of PKA (1CMK, light grey and white) with the closed form of CK2 (dark grey and white) picks out conserved regions in both the small and large lobes of the structures, (a) The large lobe alignment covers 132 of 350 residues in PKA, with an RMSD of 1.19 Å. (b) The small lobe alignment covers 46 residues in the twisted β structure of the small lobe with an RMSD of 0.99 Å.

Table 4
Common substructures of 1ATP and 1F0Q

Domain 1				Domain 2			
PKA/CK2	n	$d_{\max}$	RMSD	PKA/CK2	n	$d_{\max}$	RMSD
38–51/34–47	14	1.63	1.23	47-51/43-47	5	1.97	1.37
54–62/50–58	9	1.35	1.05	54-58/50-54	5	1.79	1.22
66–74/62–70	9	1.95	1.29				
90–95/80–88	9	1.50	1.04	90-95/80-85	6	1.39	1.06
100–109/91–100	10	1.78	0.80	102-109/93-100	8	1.52	1.20
115–124/108–117	10	1.62	1.00	117-121/110-114	5	1.39	1.24
142–149/132–139	8	1.64	1.27				
154–160/144–150	7	1.97	1.86	139-173/129-163	35	1.89	1.89
162–174/152–164	13	1.90	1.04				
178–192/169–183	15	1.84	0.97	179-191/170-182	13	1.61	0.98
198–212/191–205	15	1.92	1.21	199-213/192-206	15	1.40	0.95
219–232/213–226	14	1.88	1.15	215-229/209-223	15	1.98	1.29
				261-275/293-307	15	1.67	0.86
				294-300/321-327	7	1.70	1.34
Total Alignment	133	1.97	1.15	Total Alignment	129	1.98	1.08

The two largest domains selected by SAT have many elements in common, but differ in the inclusion of specific elements. Domain 1 includes more strands of β-sheet, while Domain 2 includes more of the helices. For each domain the first column gives the corresponding residues in the two proteins, the second column gives the size of the fragment, the third column gives the maximum distance between corresponding superposed atoms within the fragment, and the fourth column gives the root mean square deviation of all superposed pairs of atoms in the fragment.

Table 5
Common substructures of closed and open forms of PKA

Domain 1				Domain 2			
Residues	n	$d_{\max}$	RMSD	Residues	n	$d_{\max}$	RMSD
15–38	24	1.98	1.16	18–48	31	1.98	1.11
				57–63	7	1.45	1.01
67–72	6	1.71	1.31	66–129	64	1.79	1.00
88–111	24	1.84	0.94				
117–316	200	1.83	0.76	138–162	25	1.66	1.06
				171–185	15	1.67	1.22
				299–315	17	1.65	1.18
345–350	6	1.99	1.45	335–350	16	1.63	0.91
Total Alignment	260	1.99	0.86	Total Alignment	175	1.98	1.06

The open and closed forms of PKA differ primarily by an opening of the active site cleft, as shown in Fig. 5. The largest piece in Domain 1, 200 residues, has substantial overlap with Domain 2 because residues close to the axis of rotation do not move as much as residues farther from the axis. Columns are as in Table 4.

shape change. The very large core domains shown in Fig. 5 demonstrate that this motion is primarily a domain movement. This is reinforced by the data in Table 5, which gives the relevant domains determined by SAT. Domain 1 consists primarily of the large lobe, with the largest piece being 200 residues that superpose with an RMSD of 0.76 Å. Domain 2 is smaller, but still contains half of the molecule. There is substantial overlap between the two domains close to the center of motion Fig. 6.

Table 6
Common substructures of 1CMK and 1F0Q

1CMK/1F0Q	n	$d_{\max}$	RMSD
64–71/60–67	8	1.97	1.55
90–96/80–86	7	1.94	1.12
102–109/93–100	8	1.52	1.10
116–123/109–116	8	1.90	1.38
139–173/129–163	35	1.95	0.96
179–191/170–182	13	1.22	0.81
199–213/192–206	15	1.50	1.05
215–230/209–224	16	1.76	1.34
261–275/293–307	15	1.70	0.96
294–300/321–327	7	1.75	1.35
Total Alignment	132	1.97	1.12

The largest domain of the structural alignment of CK2 onto the open conformation of PKA shows fragments very similar to those from the alignment of CK2 onto the closed conformation, demonstrating that the conservation of shape can be detected in the presence of both conformational change and differences in sequence. Note that many of the pieces are very close to ones listed in Table 4.

The comparison of CK2 with the open form of PKA shows how well conserved substructures can be identified by SAT. CK2 has significant insertions and deletions with respect to PKA, and is in a different conformation than the open form of PKA. The largest domain found by SAT is shown in Table 6. It contains 132 residues and has an RMSD of 1.12 Å. It also contains many of the pieces previously identified as having the same shape when comparing the closed form of PKA and CK2, including the 35 residue piece 139–173.

5. Conclusions

A method for precise determination of conformational similarities and differences has been developed based on identification of pieces of structure with common shapes. The method is intended as a microscope for studying structural details, rather than a telescope for detection of distant similarities. The utility of the method has been demonstrated by examples of structurally similar but very distantly related proteins, by study of structural similarity between two different protein kinases, by study of conformational change of one molecule, and by comparison of two homologous structures in different conformational states. In all cases the method revealed useful information about the structures.

Acknowledgments

This research was supported by the National Science Foundation (DBI 9911196) and the National Institutes of Health (5P41RR008605 and 1P01HL48018). We thank Chino Nojo Inc. for its partial support of this work.

References

[1] W.L. Nichols, B.H. Zimm, L.F. Ten Eyck, Conformation-invariant structures of the $\alpha_1\beta_1$ human hemoglobin dimer, J. Mol. Biol. 270 (1997) 598–615.

[2] D. Knighton, J. Zheng, L.F. Ten Eyck, V. Ashford, N.-h. Xuong, S. Taylor, J. Sowadski, Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase, Science 253 (1991) 407–413.

[3] D. Knighton, J. Zheng, L. Ten Eyck, N.-h. Xuong, S. Taylor, J. Sowadski, Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase, Science 153 (1991) 414–420.

[4] I. Tsilgeny, B.D. Grant, S.S. Taylor, L.F. Ten Eyck, Catalytic subunit of cAMP-dependent protein kinase: Electrostatic features and peptide recognition, Biopolymers 39 (1996) 353–365.

[5] I. Tsigelny, J.P. Greenberg, S. Cox, W.L. Nichols, S.S. Taylor, L.F. Ten Eyck, 600 ps molecular dynamics reveals stable substructures and flexible hinge points in cAMP dependent protein kinase, Biopolymers 50 (1999) 513–524.

[6] J. Zheng, D.R. Knighton, N.H. Xuong, S.S. Taylor, J.M. Sowadski, L.F. Ten Eyck, Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations, Protein Sci. 2 (10) (1993) 1559–1573.

[7] N. Narayana, S. Cox, X. Nguyen-huu, L.F. Ten Eyck, S.S. Taylor, A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility, Structure 5 (7) (1997) 921–935.

[8] C.M. Smith, M. Gribskov, I. Shindyalov, S.S. Taylor, L.F. Ten Eyck, S. Veretnik, P.B. Bourne, The protein kinase resource, Trends Biochem. Sci. 22 (1997) 444–446.

[9] D.R. Knighton, R.B. Pearson, J.M. Sowadski, A.R. Means, L.F. Ten Eyck, S.S. Taylor, B.E. Kemp, Structural basis of the intrasteric regulation of myosin light chain kinases, Science 258 (5079) (1992) 130–135.

[10] W.L. Nichols, G.D. Rose, L.F. Ten Eyck, B.H. Zimm, Rigid domains in proteins: An algorithmic approach to their identification, Proteins: Struct. Funct. Genet. 23 (1995) 38–48.

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Databank, Nucl. Acids Res. 28 (2000) 235–242.

[12] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH—a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.

[13] I.N. Shindyalov, P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CH) of the optimal path, Protein Eng. 11 (1998) 739–747.

[14] L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, J. Mol. Biol. 233 (1993) 123–138.

[15] P. Lackner, W.A. Koppensteiner, M.J. Sippl, F.S. Domingues, ProSup: a refined tool for protein structure alignment, Protein Eng. 13 (2000) 745–752.

[16] A.G. Murzin, S.E. Brenner, H.T.C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[17] L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A. Murzin, SCOP database in : refinements accommodate structural genomics, Nucl. Acids Res. 30 (2002) (2002) 264–267.

[18] A.-S. Yang, B. Honig, An integrated approach to the analysis and modeling of protein sequences and structures. I. protein structural alignment and a quantitative measure for protein structural distance, J. Mol. Biol. 301 (2000) 665–678.

[19] T. Madej, J.-F. Gibrat, S.H. Bryant, Threading a database of protein cores, Proteins: Struct. Fund. Genet. 23 (1995) 356–369.

[20] J.-F. Gibrat, T. Madej, S.H. Bryant, Surprising similarities in structure comparison, Curr. Opinion Struct. Biol. 6 (1996) 377–385.

[21] N. Leibowitz, R. Nussinov, H.J. Wolfson, MUSTA—a general, efficient, automated method for multiple structure alignment and detection of common motifs: Application to proteins, J. Comput. Biol. 8 (2) (2001) 93–121.

[22] B. O'Neill, Elementary Differential Geometry, Academic Press, San Diego, 1997.

[23] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, Cambridge, UK, 1992.

[24] S. Rackovsky, H.A. Scheraga, Differential geometry and polymer conformation I: Comparison of protein conformations, Macromolecules 11 (1978) 1168–1174.

[25] G.S. Chirikjian, A.B. Kyatkin, Engineering Applications of Non-commutative Harmonic Analysis: With Emphasis on Rotation and Motion Groups, CRC Press, Boca Raton, Florida, 2000.

[26] D. Knuth, The Art of Computer Programming, Vol 1, Addison-Wesley, Boston, 1968.

[27] W. Kabsch, A solution for the best rotation to relate two sets of vectors, Acta Crystallog. Sect. A: Found. Crystallog. Part A 32 (1976) 922–923.

[28] W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors, Ada Crystallog. Sect. A: Found. Crystallog. Part A 34 (1978) 827–828.

[29] A.G. Murzin, Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors, J. Mol. Biol. 230 (1993) 689–694.

[30] J.R. Somoza, F. Jiang, L. Tong, C.-H. Kang, J.M. Cho, S.-H. Kim, Two crystal structures of a potently sweet protein: Natural monellin at 2.75 Å resolution and single-chain monellin at 1.7 Å resolution, J. Mol. Biol. 234 (1993) 390–404.

[31] W. Bode, R. Engh, D. Musil, U. Thiele, R. Huber, A. Karshikov, J. Brzin, J. Kos, V. Turk, The 2.0 Å X-ray crystal structure of chicken egg while cystatin and its possible mode of interaction with cysteine proteinases, EMBO J. 7 (1988) 2593–2599.

[32] P. Kraulis, MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, J. Appl. Crystallog. 24 (1991) 946–950.

[33] J. Zheng, D.R. Knighton, L.F. Ten Eyck, R. Karlsson, N. Xuong, S.S. Taylor, J.M. Sowadski, Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor, Biochemistry 32 (9) (1993) 2154–2161.

[34] R. Battistutta, S. Sarno, E. De Moliner, E. Papinutto, G. Zanotti, L.A. Pinna, The replacement of ATP by the competitive inhibitor emodin induces conformational modifications in the catalytic site of protein kinase CK2, J. Biol. Chem. 275 (2000) 29618–29622.

[35] M.G. Kendall, Rank Correlation Methods, Griffin, London, 1970.

[36] J. Hajek, Z. Sidak, Theory of Rank Tests, Academia, Prague, 1967.

[37] E.W. Myers, W. Miller, Optimal alignments in linear space, Comput. Appl. Biosci. 4 (1988) 11–17.

[38] S.K. Hanks, A.M. Quinn, T. Hunter, The protein kinase family—conserved features and deduced phylogeny of the catalytic domain, Science 241 (1988) 42–52.