BMC Bioinformatics

Methodology article

Open Access

Alignment of protein structures in the presence of domain motions

Roberto Mosca[1,3], Barbara Brannetti[1,4] and Thomas R Schneider*[1,2,3]

Address: [1]IFOM, the FIRC Institute for Molecular Oncology Foundation, Via Adamello 16, 20139, Milan, Italy, [2]European Institute of Oncology, Via Ripamonti 435, 20141, Milan, Italy, [3]European Molecular Biology Laboratory Hamburg Outstation c/o DESY, Notkestraße 85, 22607, Hamburg, Germany and [4]Novartis Pharma AG, CH-4056 Basel, Switzerland

Email: Roberto Mosca - roberto.mosca@embl-hamburg.de; Barbara Brannetti - barbara.brannetti@novartis.com; Thomas R Schneider* - thomas.schneider@embl-hamburg.de

* Corresponding author

Published: 27 August 2008

BMC Bioinformatics 2008, **9**:352 doi:10.1186/1471-2105-9-352

Received: 12 March 2008
Accepted: 27 August 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/352

Abstract

Background: Structural alignment is an important step in protein comparison. Well-established methods exist for solving this problem under the assumption that the structures under comparison are considered as rigid bodies. However, proteins are flexible entities often undergoing movements that alter the positions of domains or subdomains with respect to each other. Such movements can impede the identification of structural equivalences when rigid aligners are used.

Results: We introduce a new method called RAPIDO (Rapid Alignment of Proteins in terms of Domains) for the three-dimensional alignment of protein structures in the presence of conformational changes. The flexible aligner is coupled to a genetic algorithm for the identification of structurally conserved regions. RAPIDO is capable of aligning protein structures in the presence of large conformational changes. Structurally conserved regions are reliably detected even if they are discontinuous in sequence but continuous in space and can be used for superpositions revealing subtle differences.

Conclusion: RAPIDO is more sensitive than other flexible aligners when applied to cases of closely homologues proteins undergoing large conformational changes. When applied to a set of kinase structures it is able to detect similarities that are missed by other alignment algorithms. The algorithm is sufficiently fast to be applied to the comparison of large sets of protein structures.

Background

When comparing structures of related proteins with different amino-acid sequences it is necessary to first perform a structural alignment, i.e. to define an equivalence map between the residues in the different structures based on their relative position in space. Once structures have been successfully aligned in three dimensions, similarities and differences can be studied in order to understand function and behaviour of the molecules under consideration.

It has been demonstrated that the problem of defining an equivalence map for residues in protein structures has no unique optimal solution [1] and that it remains computationally hard [2-4] even when it is described by a well defined optimization function. Nevertheless, many tools have been created for the pairwise and the multiple alignment of protein structures using different heuristics to produce results on acceptable time-scales (for comprehensive reviews see [5-7]).

Alignment methods can be classified based on whether the two structures to be aligned are considered as rigid bodies or whether internal flexibility between domains or subdomains is accommodated in the alignment. Methods belonging to the group of 'rigid aligners' are SSAP [8], CE [9], ProSup [10], KENOBI [11], MAMMOTH [12], TOPOFIT [13], TM-align [14], SABERTOOTH [15] and TetraDA [16]. DALI [17] allows for limited molecular flexibility through the use of an elasticity term in its similarity function, but nevertheless is considered to be a rigid aligner [18]. The group of rigid aligners also includes algorithms like VAST [19] and SSM [20] that, in order to produce alignments rapidly, first identify correspondences between secondary structure elements (SSE) and then extend the alignment to the residue level. Several rigid aligners have been extended for addressing the multiple alignment problem (CE-MC [21] and MAMMOTH-Mult [22]).

As it is well known, protein molecules are flexible entities with internal movements ranging from the displacement of individual atoms to movements of entire domains or subdomains [23,24]. Large-scale movements of groups of atoms complicate the correct identification of structural equivalences between related proteins when rigid structural aligners are used.

The molecular chaperon GroEL is an interesting case of protein molecules exhibiting pronounced molecular flexibility between structurally conserved domains. By comparison of crystal structures of different functional states, the GroEL molecule can be divided into three domains (equatorial, hinge and apical) separated by hinge regions [25]. Due to the large relative motion of the domains between different functional states, rigid body aligners will typically fail to align crystal structures of GroEL with different sequences in different conformational states.

In recent years, tools for the flexible alignment of protein structures have been introduced. These tools find an equivalence map between the residues of two molecular structures even when substantial intramolecular movements occur around molecular hinges. The regions between hinge points are commonly considered as rigid bodies and the alignment is usually optimized to minimize the number of hinges. The group of 'flexible aligners' includes, FlexProt [26] and FATCAT [18] and their corresponding extensions to multiple alignment MultiProt [27] and POSA [28].

However, in alignments of molecules such as GroEL where the polypeptide chain folds back onto itself (Figure 1) and thereby creates structural domains in which parts of the polypeptide chain that are distant in sequence space engage into stable contact in three-dimensional space (e.g. for the equatorial domain of GroEL, see below), many of these aligners meet difficulties in recognizing the spatial continuity as will be illustrated below.

Here we introduce a new algorithm for the flexible structural alignment of proteins called RAPIDO (for Rapid Alignment of Proteins in terms of Domains). RAPIDO is capable of aligning related protein molecules in the presence of large conformational differences while at the same time groups of equivalent parts of the polypeptide that are distant in sequence but nevertheless form spatially continuous domains are identified correctly as such. As a first step RAPIDO creates an equivalence map between the two structures by taking into account flexibility, with a procedure that is similar to the one used by FATCAT [18]. This step is followed by the application of a genetic algorithm [29] for the identification of *structurally conserved regions* that can be continuous in space but not in sequence (e.g. the equatorial domain of GroEL). The result of the procedure is a description of a protein in terms of structurally conserved regions connected by localized hinges or by flexible linker regions. We have chosen the standard parameter settings for RAPIDO such that more emphasis is placed on the geometric similarity of the structurally conserved regions (as reflected in low RMSDs) than on their size (as reflected in the length of the alignments). With this choice, the resulting structurally conserved regions will have a high level of similarity allowing their usage for robust coordinate-based structure superpositions.

In the following, we describe the algorithm used and the application of RAPIDO to a number of test cases. For all test cases, RAPIDO produces results that are in agreement with previous analyses. Regions identified as structurally conserved furnish subsets of atoms whose relative positions between different structures are very well maintained. Superpositions based on these subsets of atoms are particularly revealing when molecular flexibility is studied.

Results

Algorithm

The alignment algorithm consists of four steps:

1. Search of short structurally similar fragments in pairs of structures, so called Matching Fragment Pairs (MFPs)

2. Chaining of the MFPs through a graph-based algorithm

3. Refinement of the alignment

4. Identification of rigid bodies

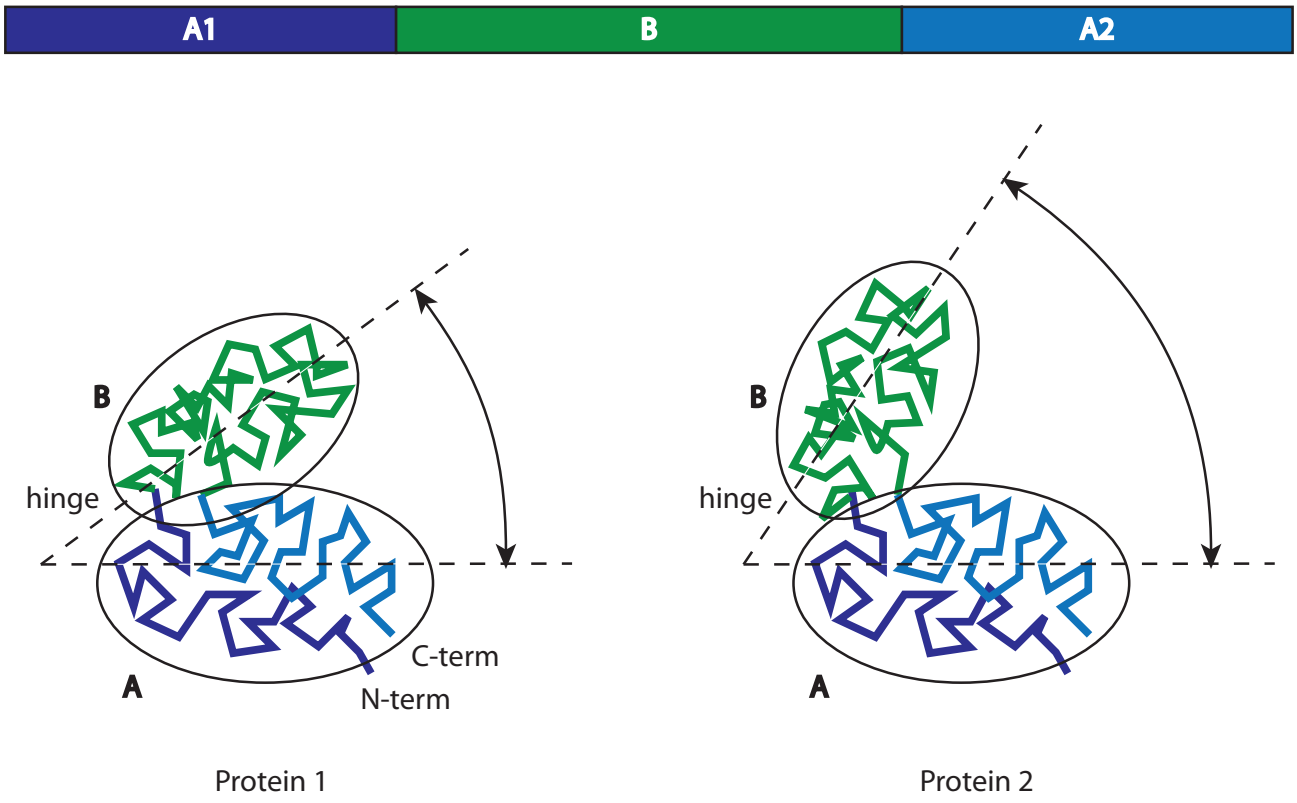

Figure 1
Alignment of two proteins with a conformational change and a polypeptide chain folding back onto itself. For two hypothetical proteins with homologous structures (protein 1 and protein 2), with two domains (one consisting of stretches A1 and A2 and one consisting of stretch B in sequence space) moving with respect to one another around a hinge, the aligned parts of the sequence are shown at the top, while the mapping of the alignment onto structures is shown with the same colours in the bottom of the figure. The alignment of proteins of such topology (e.g. GroEL) poses two problems: (1) the treatment of large conformational changes involving the motion of domains around hinge-regions (closed form of protein 1 versus open form of protein 2) and (2) the recognition of domains that are continuous in space but discontinuous in sequence (domain A of protein 1 and protein 2 consisting of parts of the N- (A1) and C-termini (A2)).

In the remainder of this section we will refer to two structures being compared as structures A and B. The *i*-th residue in structure X (X = A or X = B) is represented by the coordinates of its C_α atom and will be indicated by $\mathbf{x}_i$ ($\mathbf{a}_i$ and $\mathbf{b}_i$ respectively).

Finding matching fragments

We define a *fragment* as an ungapped stretch of residues and a *matching fragments pair* (MFP) as a pair of structurally similar fragments of the same length in two structures being compared. The search for MFPs is in fact implemented in a number of alignment tools as the initial step [9,10,18,26,30] because it significantly reduces the complexity of the search space for the alignment. Pairs of similar fragments named *matching fragment pairs* (MFPs) here, have been named *aligned fragment pairs* (AFPs) in other publications [9,18,30]. In the context of the RAPIDO aligner, we prefer to use the notation of *matching fragment pairs* in order to clarify that in a later stage of the alignment algorithm, a subset of the *matching fragment pairs* forming the initial set is selected to assemble the actual alignment, and the selected MFPs thus become *aligned fragment pairs*.

While many algorithms use the RMSD to measure the similarity between two fragments [18,26,30], we use an alternative measure, the sum of the absolute values of the elements of the difference distance matrix between the C_α-atoms of the two fragments (eq. 1 in the *Methods* section).

At first an exhaustive search for MFPs of length m_L (m_L = 8 in the implementation) is performed, followed by a clustering step in which overlapping MFPs are joined to form longer ones.

Chaining matching fragment pairs and refining the alignment

The MFPs identified in the first step constitute a set of potential building blocks for the final alignment from which, in the second step, a subset of MFPs representing a structural alignment is assembled. This is done by casting the problem into a graph representation to which a standard algorithm for identification of the longest path is applied. The MFPs are represented as vertices of a graph and two MFPs (e.g. two vertices) are connected by an edge if they are topologically ordered, i.e. if they are composed of two pairs of fragments that appear in the same order in the two residue sequences. Every path in the graph represents a possible alignment and by choosing an appropriate *weight function* for the edges, the problem of finding the best alignment is translated into the problem of identifying the longest path on a graph. We solve this problem by applying a dynamic programming algorithm for the identification of the longest path. The alignment obtained in this way is a preliminary alignment that is then refined (details on the refinement process can be found in the *Methods* section) resulting in the *raw alignment*.

Identification of rigid bodies and flexible superposition

Once the raw alignment has been calculated, the algorithm performs a search for structurally conserved regions. Structurally conserved regions relate to conformationally invariant regions detected in different conformations of the same molecule as described in [31]. Conformationally invariant regions can be defined as subsets of equivalent atoms whose interatomic distances are identical within error between the different conformations of the same molecule [31]. In the comparison of different molecules, the concept can be generalized by considering subsets of *aligned* residues for which distances between C_α-atoms are identical within a tolerance as *structurally conserved regions*. These subsets can be identified using a genetic algorithm operating on scaled difference distance matrices [29,31,32]. In our previous work the elements of the difference distance matrix were scaled by propagated coordinate errors resulting in error-scaled difference distance matrices [31]. The parameters necessary for the estimation of the coordinate errors were extracted automatically from the PDB files and if necessary corrected manually. This approach is not applicable when very many PDB-files are being investigated in the context of searching for related structures in large data bases as the values extracted can be unreliable mostly caused by human errors made when the parameters where entered in the first place. For the purpose of structural alignment, we therefore use a simplified approach in which the estimate for the coordinate error of an atom i with a B-value of B_i is replaced by an analogous quantity $\tilde{\sigma}_i$ calculated as follows

$$\tilde{\sigma}_i = k \cdot \left(1 + \frac{B_i}{2\pi^2} \right)^{\eta},$$

where the constants k and η have been empirically optimized to 0.4 Å and 2/3. $\tilde{\sigma}_i$ can then be propagated into a scaling-factor for difference distance matrix elements in a manner similar to the previous treatment.

The algorithm searches iteratively for structurally conserved regions in analogy to the approach presented in [32]). Aligned residues that cannot be assigned to structurally conserved regions are marked as flexible.

To characterize the agreement between two structures after the equivalent residues have been divided into structurally conserved and flexible regions, separate least-square superpositions are performed for the different structurally conserved regions.

Based on this superposition allowing flexibility between conserved parts of a three-dimensional structure, we define the 'flexible RMSD' ($RMSD_f$) as the standard RMSD calculated for all pairs of equivalent C_α-atoms after separate least-squares superposition for the different structurally conserved regions.

Testing

In order to assess the functionality of the method, we applied it to various test cases. Here, we describe the analysis of two structures of different topologies with known hinge-motions (Ran and GROEL) and we compare the results of RAPIDO with those obtained by FATCAT [18] and FlexProt [26]. Second, we compare the results obtained with RAPIDO with those given by DALI for 2278 pairwise alignments between 68 crystal structures of protein kinases from human.

Ran

Ran is a small GTPase belonging to the Ras superfamily that plays an important role in several nuclear functions, including nucleocytoplasmic transport, cell-cycle progression and nuclear envelope assembly [33]. Here we compare two structures of Ran proteins from two different organisms: the first one is the structure of a Q69L mutant of Ran from dog with a bound GDP molecule (RanQ69L*GDP, PDB id 1byu, [33]); the second structure corresponds to Ran from human in complex with human RanBP2 and a non-hydrolysable GTP analogue (Ran*GppNHp complex, PDB id 1rrp, [34]).

The RAPIDO alignment shows that major parts of the two structures are very similar. 182 residues are aligned, 158 of which are assigned to two rigid bodies. The first rigid body covers more than 70% (140 residues) of the entire protein, can be superposed with an RMSD of 0.76 Å (Figure 2) and corresponds to the main body of the protein. Two fragments in this region are either not aligned or aligned but marked as flexible. They correspond to the well-known SWITCH I and SWITCH II regions, which exhibit different conformations depending on the type of bound nucleotide and regulate the interactions of the protein with nuclear trafficking components [35]. The C-terminal regions of the two structures have been aligned although they are located in very different positions with respect to the main body of the protein in the two structures. This region is composed of an unstructured loop followed by a helix that is known to assume a different conformation depending to the GTP/GDP-binding state of the protein [36]. The C-terminal helix is attached to the main body of the protein in the Ran*GDP complex while in the Ran*GppNHp complex, it interacts with a groove on the surface of the RanBD1-domain approximately 25 Å distant from the Ran main body. While the helix is recognized as a second rigid body, the part of Ran connecting its main body with the C-terminal helix in different conformations is marked as a flexible region.

The alignments between the two structures as produced by FATCAT and FlexProt are slightly longer (186 aligned residues for FATCAT, 188 for FlexProt). The separation between the two rigid bodies is similar in the three alignments but the RMSD for the superposition of the single rigid regions is higher in FATCAT and FlexProt alignments than in the RAPIDO alignment. This is due to the fact that in these two aligners all aligned residues are used for the superposition while RAPIDO distinguishes between structurally conserved and flexible aligned residues and uses only the residues in structurally conserved regions to perform the superposition. In fact, in the FATCAT and FlexProt align fragments the SWITCH I and II loops are attributed to the first equivalent region yielding an RMSD for the superposition of this first rigid part of 1.51 Å for FATCAT and 2.87 Å for FlexProt. The unstructured loop connecting the main body and the C-terminal helix is partly assigned to the first equivalent region and partly to the second adding to the increased RMSD-values for the respective superpositions.

Although, for this case, the alignments are mostly equivalent, the one provided by RAPIDO highlights the different conformations of three important functional elements corresponding to the SWITCH I and II loops and to the C-terminal loop and produces an accurate superposition of the two structures in which these differences can be clearly analyzed.

GroEL

GroEL is a bacterial chaperonin that, together with its co-chaperonin GroES forms a system helping newly synthesized polypeptides to reach their native state in the crowded cellular environment. GroEL consists of 14 identical subunits that are assembled as two heptameric rings stacked back to back, forming a cavity in the centre in which a newly formed polypeptide can find a protected environment for refolding [37]. Each subunit corresponds to a single protein molecule with three domains called the equatorial, apical and hinge domain (Figure 3b). During its activity, the GroEL complex undergoes dramatic conformational changes correlated with different relative arrangements of the three domains in each subunit. Here we align the structure of one GroEL subunit from *Escherichia coli* (PDB id 1OEL, [38]) with one from *Thermus termophilus* in complex with ADP (PDB id 1WE3, [39]).

The structural alignment produced by RAPIDO covers 98% of the molecule (516 aligned residues), with a flexible RMSD of 0.88 Å. Four structurally conserved regions are identified (Figures 3b and 3e) corresponding to the three canonical domains of the GroEL subunit plus the stem loop in the equatorial domain comprising approximately 20 residues. The three structurally conserved regions are in different relative positions with respect to each other in the two structures as highlighted by the RMSD of 11.59 Å for the rigid superposition. However, by examining the superposition of the structurally conserved regions separately, the structural conservation of major parts of GroEL can be well appreciated both from the RMSDs ranging between 0.81 and 1.04 Å and the actual superposition (Figure 3). In addition to the three large canonical domains, the so-called stem loop in the equatorial domain is found to constitute a small structurally conserved region assuming different orientations in the two structures. This dependence of the positions of the stem loop on the functional state had already been observed by Xu et al. [25].

The alignment produced by FATCAT has approximately the same length (518 residues) and a flexible RMSD of 2.45 Å. Two hinges are identified and the structure is divided into the three regions shown in Figure 3d. While the apical domain is identified by both RAPIDO and FATCAT as an equivalent region, the equivalent regions for the other two domains display marked differences. The hinge domain is, in the FATCAT alignment, joined to the equatorial domain and the resulting superposition is thus an average between the superposition of the two single subunits, leading to a higher value for the RMSD. Due to the sequential constraint imposed by FATCAT (two regions that are not sequential cannot belong to the same rigid body), the block corresponding to the equatorial-

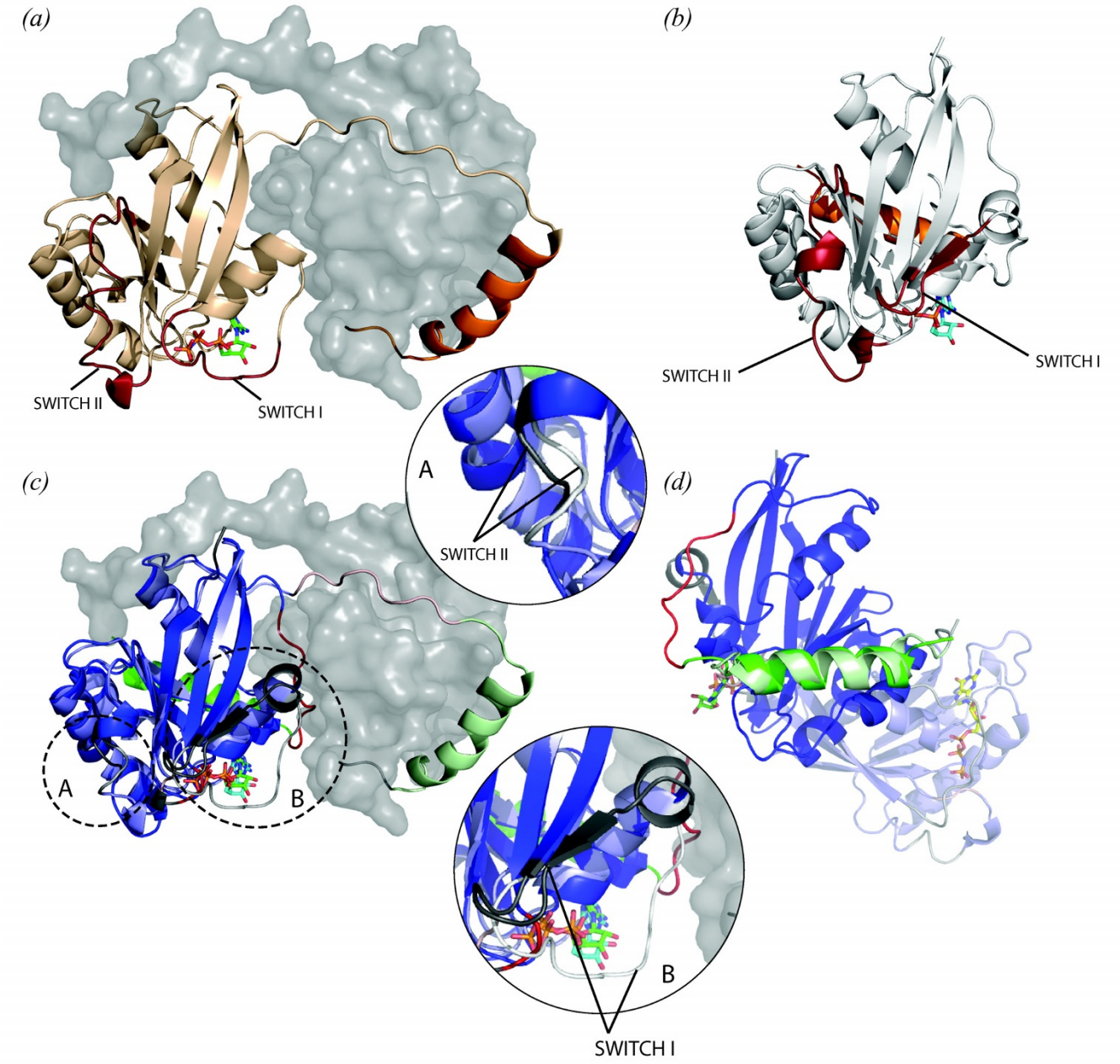

Figure 2

Alignment of structures of two Ran proteins. (a) Structure of human Ran (cartoon) bound to a non-hydrolysable GTP analogue (sticks) with the Ran-binding domain of human RanBP2 (grey surface). The SWITCH I and II loops are shown in red, the C-terminal helix is displayed in orange. (b) Structure of a Q69L mutant of canine Ran (cartoon) with a bound GDP molecule (sticks) (c) Superposition of the two Ran molecules on the first rigid body identified by RAPIDO (140 atoms, RMSD 0.76 Å). The different conformations of the SWITCH I and II fragments as well as the large displacement of the C-terminal helix are clearly visible. In this figure (and in all other figures), the first rigid body is colored in blue, the second in green, the third in cyan, the fourth in magenta. Parts of structures that cannot be aligned are marked in grey. Parts of structures that were aligned and then identified as having different conformations in different structures are colored red. When two structures are compared, one is shown in light, the other in dark colors – here the structure of the protein from human is shown in dark colors, while the structure from dog is shown in light colors. (d) Superposition of the two Ran molecules on the second rigid body consisting mostly of the C-terminal helix (in green, 18 atoms, RMSD 1.35 Å). The unstructured linker preceding the C-terminal helix has been found to be flexible and is marked red. All figures were produced with PyMOL http://pymol.sourceforge.net/.

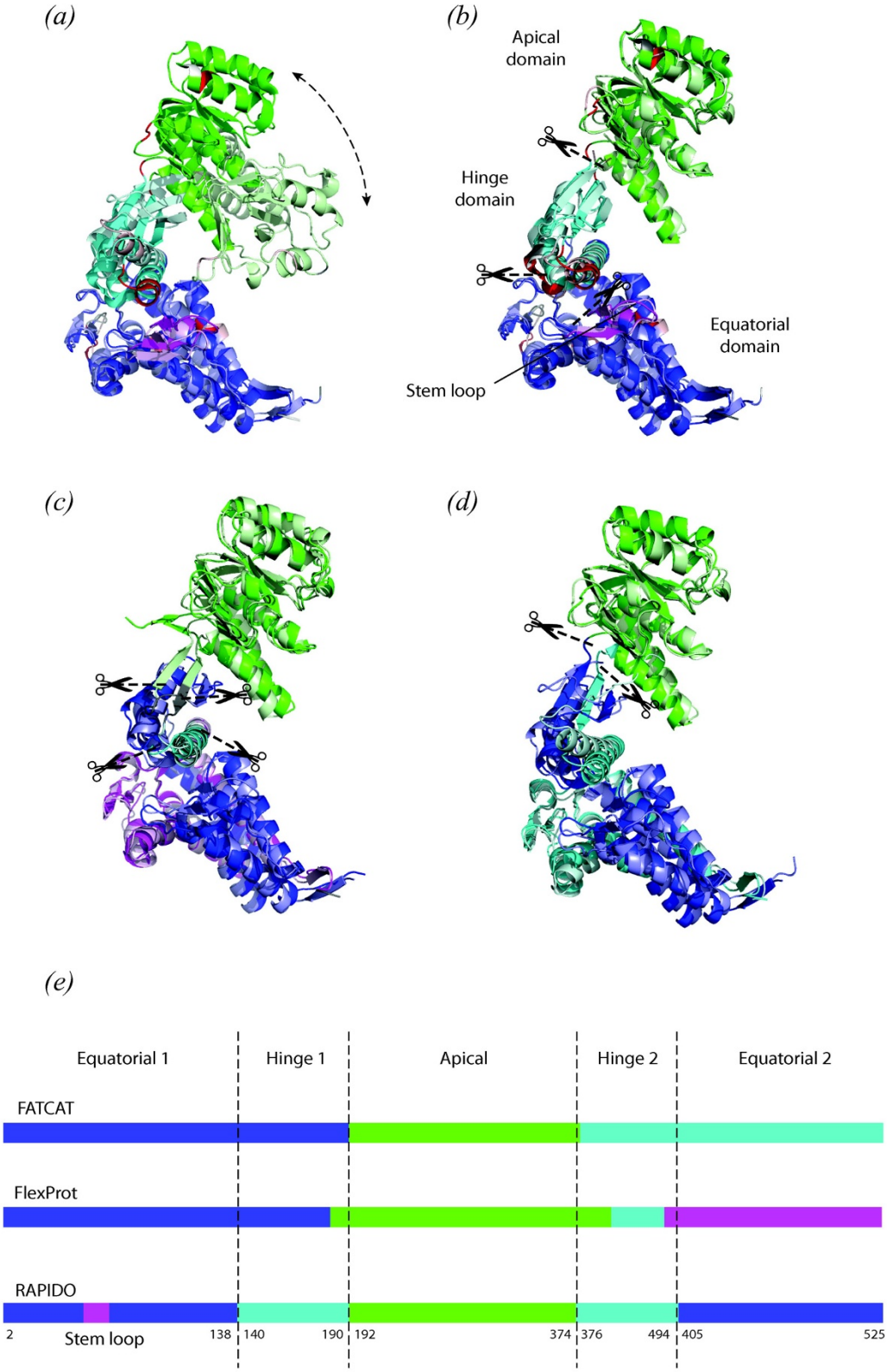
(a)
(b)
Apical domain
Hinge domain
Equatorial domain
Stem loop
(c)
(d)
(e)
Equatorial 1
Hinge 1
Apical
Hinge 2
Equatorial 2
FATCAT
FlexProt
RAPIDO
2
Stem loop
138
140
190
192
374
376
494
405
525

Figure 3

Figure 3
Alignment of two structures of GroEL from *Thermus Thermophilus* (1we3) and *Escherichia Coli* (1oel). (a) Superposition of the two structures on the first rigid body identified by RAPIDO (in blue, 1we3 is in darker colors while 1oel is in lighter colors). (b) Flexible superposition based on the rigid bodies identified by RAPIDO. Scissor symbols indicate the points in which the 1oel was divided in order to separately superpose the regions identified as rigid bodies (1st rigid body: 220 atoms, RMSD 0.81 Å; 2nd rigid body: 178 atoms, RMSD 0.93 Å; 3rd rigid body: 71 atoms, RMSD 1.04 Å; 4th rigid body: 20 atoms, RMSD 0.68 Å). (c) Flexible superposition generated by FlexProt (1st fragment: 122 atoms, RMSD 2.62 Å; 2nd fragment: 21 atoms, RMSD 3.02 Å; 3rd fragment: 193 atoms, RMSD 2.95 Å; 4th fragment: 177 atoms, RMSD 2.95 Å). (d) Flexible superposition generated by FATCAT (1st fragment: 186 atoms, RMSD 3.17 Å; 2nd fragment: 179 atoms, RMSD 0.96 Å; 3rd fragment: 153 atoms, RMSD 3.17 Å). (e) Mapping of the conserved domains identified by different methods onto the primary sequence. Residue numbers of domain boundaries in the *E. Coli* structure (1oel) as determined by RAPIDO are indicated; small flexible insertions within the domains have been left out for clarity.

hinge domain is split into two fragments corresponding to the N- and C-terminal parts. The stem loop is in the FATCAT alignment included in the first rigid region.

FlexProt creates an alignment of 513 residues with a flexible RMSD of 2.87 Å. Three hinge-points dividing the structure in four fragments are identified. As in the FATCAT alignment, the apical domain is kept separate from the rest of the structure. Even if the C-terminal parts of the hinge and equatorial domains are separated by a hinge point, their N-terminal counterparts are kept together including the stem loop. In general, the alignments produced by FATCAT and FlexProt tend to underestimate the number of hinges for this pair of structures and cannot be used to highlight the difference between the equatorial and hinge domains, nor the different conformation of the stem loop.

A correct delineation of the domains is of particular interest in this case. In fact, the identified domains can be used as rigid bodies for the interpretation of low-resolution electron density maps for GroEL in different functional states as determined by electron microscopy. In this way, they allow to derive conclusions at the atomic level from lower resolution data (e. g. Ranson et al. [40]).

Human kinase structures

Protein kinases are multi-domain proteins catalyzing the phosphorylation of proteins and play important roles in controlling many cellular processes (chapter 13 in [41]). The protein kinase catalytic domain consists of two lobes, a small N-terminal lobe and a large C-terminal lobe connected by a hinge region and is often augmented by other domains that serve in regulation of the kinase activity. Prominent examples of such domains are the SH2 and SH3 domains present in protein kinases such as src Hck kinase [42] and Bcr-abl kinase [43]. In protein kinases, the relative positions and orientations of the different domains are very variable and depend on many factors such as the binding of ligands in the active site and/or the presence of regulating factors.

We used RAPIDO to perform an all-against-all alignment for 68 structures of human protein kinases (2278 alignments in total). For comparison, for every pair of structures, an alignment was also determined using DaliLite Ver. 2.4.4 (the standalone version of DALI).

Alignments produced by RAPIDO and DALI are compared in Figure 4 and summarized in additional file 1. In terms of overall length, the majority of the alignments are comparable. However, for some cases, the RAPIDO alignments are substantially longer than the DALI alignments (blue and red dots in Figure 4).

Three of these cases (blue dots in Figure 4a) correspond to alignments between the structures of Hck from Human (1AD5, [42]), c-Src from Human (1FMK, [44]), Csk from Rat (1K9A, [45]) and c-Abl from Mouse (1OPK, [43]). In these four structures, the kinase domain was crystallized in the presence of SH2 and SH3 domains. Depending on the functional state of the kinase, the SH2 and SH3 domains can be in substantially different positions with respect to the kinase domain. Such different positions can cause rigid aligners not to recognize all domains as similar. For the case of the alignments between Hck and Csk, and between c-Src and Csk (dots in the red circle in Figure 4a), DALI aligns only 329 and 350 residues respectively with the aligned residues being located in the protein kinase domain and in the SH2 domain. The SH3 domain is not included in the alignment. For the alignment between Csk and c-Abl (dot in the green circle in Figure 4a) DALI aligns only the protein kinase domain. The alignment produced by RAPIDO in all three cases is longer (389 to 399 residues) and comprises the kinase domain as well as the SH2 and SH3 domains (Figure 5).

To illustrate different positions of domains in protein kinase structures, Figure 5 shows the alignment between the structures of Hck (PDB id 1ad5) and Csk (PDB id 1k9a). Although the positions and orientations of the SH2 and SH3 domains with respect to the protein kinase domain are substantially different in the two structures

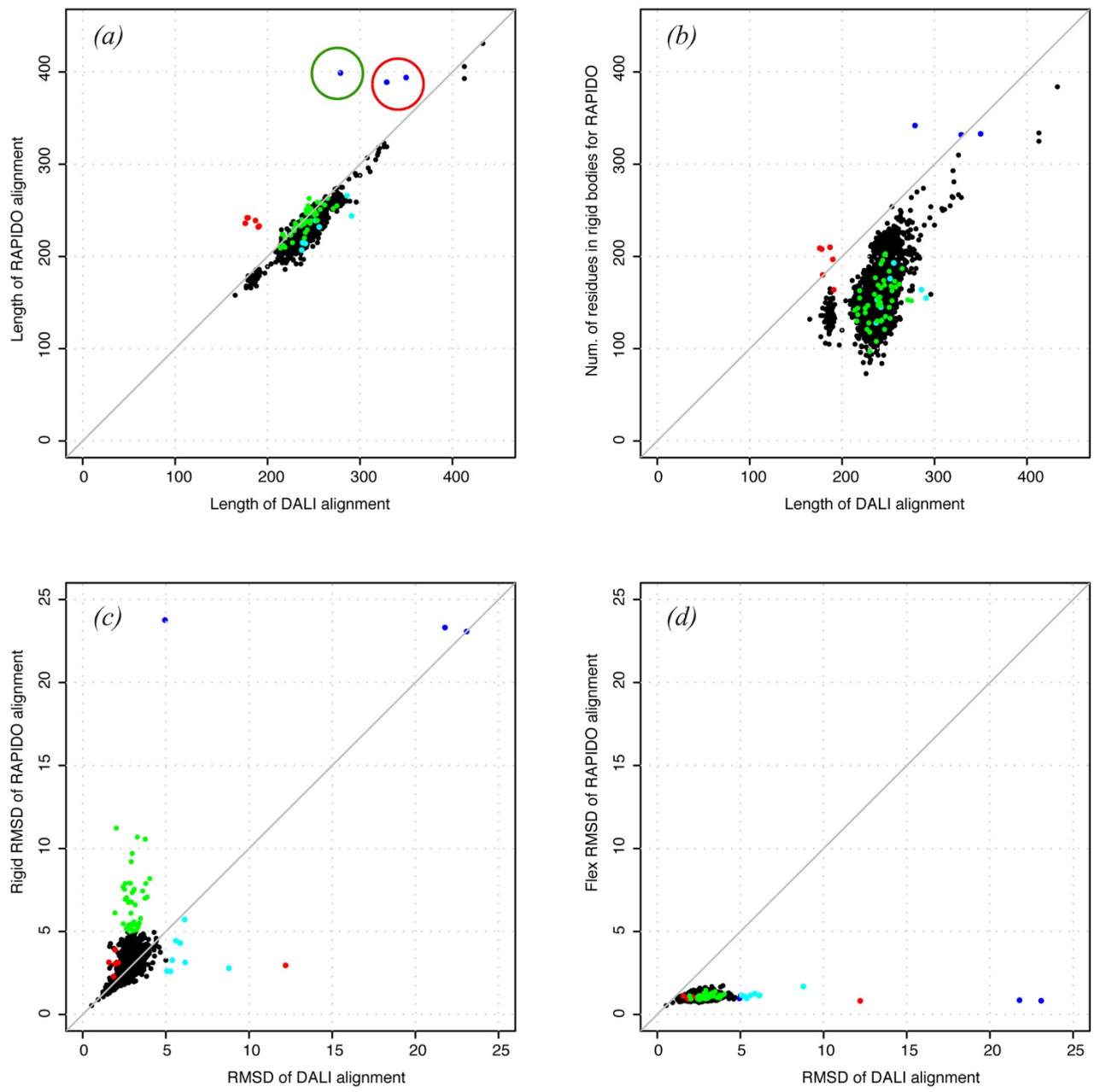

Figure 4

Comparison between DALI and RAPIDO on the dataset of human kinase structures. Every dot in the scatter plots represents one of the 2278 alignments between 68 structures (a) Length of the raw alignment provided by RAPIDO vs. the length of the corresponding DALI-alignment. Blue and red dots represent pairs of structures for which the RAPIDO alignment is significantly longer than the DALI alignment. Green and cyan dots indicated structures for which the rigid RMSD of the RAPIDO-alignment is substantially higher than that for the DALI-alignment or *vice versa* (Panel (c)). Data points surrounded by circles are discussed in the text. (b) Total number of residues assigned to rigid domains by RAPIDO vs. length of DALI alignment (c) Rigid RMSD for all atoms aligned by RAPIDO vs. rigid RMSD for atoms aligned by DALI. (d) Flexible RMSD for atom aligned and identified as belonging to rigid bodies by RAPIDO vs. rigid RMSD for all DALI-aligned atoms. Please note that the lengths and RMSDs given for the RAPIDO alignments correspond to *aligned* residues in Panels (a) and (c) while they correspond to *rigid* or *structurally conserved* residues in Panels (b) and (d); the difference between the two sets are *flexible* residues that have been aligned but are found in different conformations in the structures being compared.

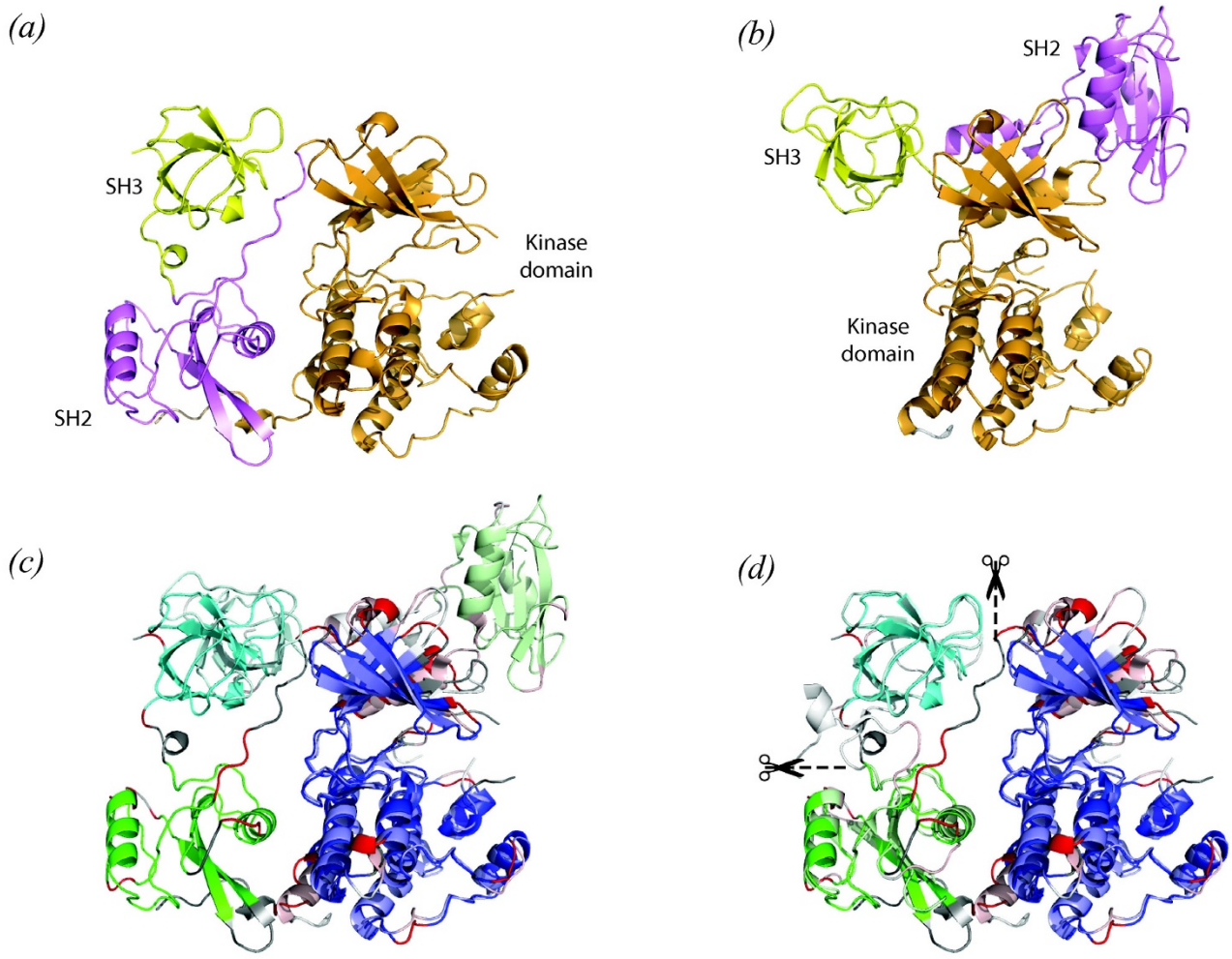

Figure 5
Alignment of structures of Hck and Csk protein kinases. Panel (a) and (b) show schematic drawings of the structures of Hck (PDB id 1ad5) and Csk (PDB id 1k9a) src kinases. The kinase domains, the SH2, and the SH3 domains are shown in orange, magenta, and yellow, respectively. (c) Superposition of both structures on the first rigid body, corresponding to the kinase domain (shown in in blue, 190 res, RMSD 0.90 Å). Hck kinase is shown in dark colors, Csk kinase in light colors. The substantially different positions of the SH2 and SH3 domains with respect to the kinase domain become visible. (d) Flexible superposition between the two structures. When superposed separately the three domains reveal a considerable level of structural conservation (1st rigid body: 190 atoms, RMSD 0.90 Å; 2nd rigid body: 81 atoms, RMSD 0.88 Å; 3rd rigid body: 55 atoms, RMSD 1.06 Å).

(Figures 5a and 5b), RAPIDO manages to align the two structures for almost their entire length identifying three separate structurally conserved regions (Figure 5c). The largest structurally conserved region corresponds to the conserved core of the protein kinase domain, while the two smaller structurally conserved regions are the SH2 and the SH3 domain. Superposition on the conserved part of the protein kinase domain clearly reveals the different positions and orientations of the SH2 and SH3 regulatory domains with respect to the catalytic domain (Figure 5c). By superposing the three regions separately (Figure 5d) the structural conservation of the different domains in the two protein structures becomes clear and a flexible $RMSD_f$ of 0.86 Å on 332 residues indicates the close relation between equivalent domains in different protein.

Other cases for which the RAPIDO alignment assigns more equivalent atoms than the DALI alignment concern alignments of structures with large differences in the opening angles measured between the N- and the C-termi-

nal lobe of the kinase domain (red dots in Figure 4). For the alignment between the structures of the protein kinase domains of CDK6 (1BI7), MAPK P38 (1P38), Src (1FMK), IGF1 receptor (1JQH), EGFR (1M17), HGFR (1R0P) and JAK3 (1YVJ), the algorithm implemented in DALI can cope with many cases of different relative domain orientation. However for the cases marked in Figure 4, the residues in the N-terminal domain are not aligned due to the large differences in opening angle between the lobes. In one of these cases (red point in Figure 4c, corresponding to the alignment between 1FMK and 1R0P), parts of the small lobe are in fact included in the alignment but at the cost of a very large RMSD between the equivalent atoms (12.20 Å for 190 atoms, Figures 4c and 4d). In all these cases, RAPIDO correctly determines the equivalences between atoms both for the C- and the N-terminal lobe, independently of their relative positions.

There are cases where the 'rigid RMSDs' measured for the superposition based on all atoms aligned by RAPIDO is substantially higher than the rigid RMSD for the corresponding DALI alignments although the alignments are of comparable length (green dots in Figures 4a and 4c). These are cases where taking into account flexibility in the RAPIDO algorithm results in an alignment including small fragments that are locally very similar but structurally not equivalent when their surrounding environment is considered. A typical situation of this kind is the erroneous alignment of periodical structural elements such as α-helices or β-strands with a shift in register. Such fragments are included in an alignment because they exhibit high local similarity and their different positions with respect to neighbouring structural elements is assumed to be due to conformational change. Although for the majority of cases, these situations are remedied, it is generally not possible to avoid them without an unacceptable loss in sensitivity. However, such incorrectly aligned fragments will not be included into structurally conserved regions as their positions in different structures are inconsistent and therefore such fragments will be marked as *aligned* but *flexible* – this is the reason for the number of residues assigned to rigid bodies by RAPIDO being usually smaller than the number of residues aligned by DALI (Figure 4b). When the flexible RMSD is calculated for all atoms assigned to structurally conserved regions (leaving out the aligned but flexible atoms), it is substantially lower than the standard RMSD calculated for the corresponding DALI alignments (Figure 4d) thus indicating the presence of similarities more clearly.

Finally, in some cases the alignment produced by DALI is longer than the one produced by RAPIDO (cyan dots in Figure 4). However, careful analysis reveals that in these cases, the DALI-alignments comprise some small fragments that are locally similar but when put in the context of their structural neighbours should actually not be considered as equivalent. The presence of such inconsistencies is also reflected in the higher values for the rigid RMSD when compared to the RAPIDO alignments (Figure 4c).

Implementation

The algorithm has been implemented in C++. For academic use, executables for various platforms can be obtained from the corresponding author upon request. A web server for aligning structures using the RAPIDO-algorithm is available at http://webapps.embl-hamburg.de/rapido.

Typical execution times with the inclusion of the pre-processing step (see *Methods* section for details) range from 0.5 sec to 1.5 s for pairs of structures between 200 and 400 residues. Without pre-processing, execution time ranges between 1.5 and 4 s for the same structures (CPU-times for iMac with an Intel Core 2 Duo processor at 2.4 GHz and 2 GB of memory running under MAC OS X version 10.4).

On output, the program generates different files. A textual representation of the alignment is generated in an HTML file. Different types of superpositions are available: rigid superposition on all aligned atoms, superpositions on individual rigid bodies and flexible superposition. The latter is obtained by subdividing the structures into pieces centred on the rigid bodies identified in the alignment procedure. The parts of the structures falling between the boundaries of two rigid bodies are moved together with the rigid body closest in sequence during the superposition.

The superposed structures with their modified coordinates are stored as PDB files. PyMOL- and RasMOL-scripts for displaying the superposed structures are generated by the program. All output information is consistently color-coded based on the rigid body assignments so that conformationally invariant parts can be easily inspected.

Conclusion

In this paper, we have introduced a new method named RAPIDO for the alignment of proteins in the presence of conformational changes. Aligned residues are grouped into subsets that can be considered as rigid domains with respect to the structures being compared; aligned residues not assigned to a rigid domain are considered flexible.

When applied to structures with known hinge motions, RAPIDO produces results that are consistent with manual analyses presented in the literature. By using a genetic algorithm operating on scaled difference distance matrices [29], structurally conserved regions are assembled con-

sistently even when composed of fragments that are not continuous with respect to the polypeptide chain.

With standard settings, RAPIDO identifies subsets of residues whose C_α-atoms can be superimposed with RMSDs of typically less than 1 Å for structurally conserved regions. Given the tight conditions in terms of similarity, the individual structurally conserved regions are generally smaller than those obtained by other alignment algorithms. However, as other regions that are in different relative positions in the structures under comparison will be aligned with high accuracy as part of different rigid bodies, the overall length of the combined alignment taking flexibility into account will be increased in many cases.

In the context of structure comparison and analysis, superpositions of structures based on atoms located in rigid domains can highlight conformational differences that, when superpositions are based on atoms sets accidentally containing flexible regions, can be difficult to identify.

The application of RAPIDO to a dataset of kinase structures showed how allowing for flexibility can help to detect similarities that are not found by rigid aligners.

To evaluate the limits of RAPIDO, we have applied the algorithm to ten 'difficult cases' of low sequence and structural similarity from Fischer's [46] dataset for benchmarking fold-recognition methods. The results obtained [see Additional file 2] indicate that for distantly related structures RAPIDO alignments are generally shorter and exhibit larger RMSDs than alignments produced by other algorithms. RAPIDO should therefore be used preferentially for cases were closely related structures are sought for.

A definite advantage of RAPIDO is the short time required to calculate an alignment. E.g., a total of 2278 alignments on a set of 68 kinase structures was completed by RAPIDO in 61 minutes. This allows applying the method presented to problems of substantial size such as querying a large set of structures for similarities with a structure of interest or all-against-all alignments of entries in structural databases.

Methods

Identification of matching fragment pairs

An MFP composed of two stretches of residues of length L starting at residue i in structure A and at residue j at structure B, is described by a triplet (i, j, L). A distance between the two fragments, $S(i, j, L)$, is calculated as:

$$S(i,j,L) = \frac{1}{L}\sum_{t=1}^{L-1}\sum_{s=0}^{t-1}\left| d_A(i+t, i+s) - d_B(j+t, j+s)\right|$$

where

$$d_x(u, v) = ||\mathbf{x}_u - \mathbf{x}_v||$$

is the element of the distance matrix between the C_α-atoms of residues u and v in structure X.

In the first step, the algorithm builds the list S^* of MFPs of length greater than or equal to m_L for which $S(i, j, L)$ is lower than a threshold m_S. Even if the number of possible fragments, $\sum_{L=m_L}^{M\equiv\min\{m,n\}}(n-L+1)(m-L+1) = O(M^3)$, is polynomial in M, finding the complete set S^* is computationally too expensive. To reduce the complexity of this step, we thus first search for all MFPs of fixed length m_L and distance $S(i, j, L)$ lower than a threshold m_S. Then we identify groups of overlapping MFPs and test whether groups of MFPs can be merged into one larger MFP. If the score for the merged MFP is lower than the chosen threshold m_S, it is kept. Technically, the merging step consists of extending a randomly chosen MFP downstream with overlapping MFPs until the score of the merged MFP becomes greater than the threshold m_S. In the current implementation of the algorithm, $m_L = 8$ and $m_S = 3.0$.

Chaining of matching fragments

In order to select the MFPs forming the alignment, first a graph representing all the MFPs identified in the first step of the algorithm is built. Every MFP becomes a vertex in the graph and two MFPs F_1 and F_2 are connected by an edge if they can be chained, i.e. if and only if $F_1 << F_2$ according to the following definition (Figure 6):

Let $F_1 \equiv (i_1, j_1, L_1)$ *and* $F_2 \equiv (i_2, j_2, L_2)$ *be two MFPs. Then,* $F_1 << F_2$ (F_1 *is less than* F_2) *if and only if*

$$((0 < i_2 - i_1 < L_1) \wedge (i_2 - i_1 = j_2 - j_1)) \vee ((i_2 - i_1 > L_1) \wedge (j_2 - j_1 > L_1))$$

This is a partial order relation and it can be demonstrated that the graph induced by the previous relation is a Directed Acyclic Graph (DAG). This graph can be formally described by the couple (V, E) where V is the set of vertices and E is the set of edges of the graph:

$$G \equiv (V, E) \quad V = \{F \equiv (i, j, L) \mid F \text{ is an MFP}\} \quad E = \{(F_1, F_2) \mid F_1 << F_2\}.$$

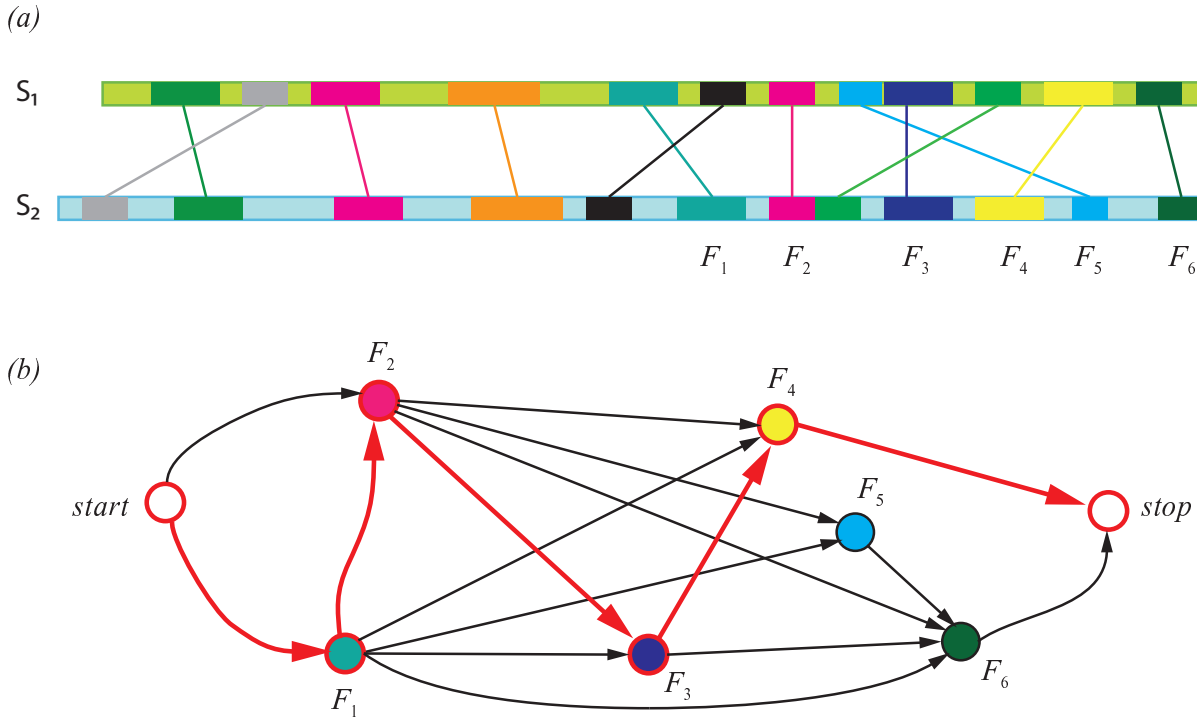

Figure 6
Chaining of Matching Fragment Pairs. A schematic representation of MFPs for two proteins with sequences S_1 and S_2. MFPs are indicated as pairs of rectangles connected by a line mapped onto the sequence in panel (a) and as nodes of a graph in corresponding colors in (b). The graph representation encodes the topological relations between the MFPs. E.g. F_3 can be chained with F_6 but it cannot be chained with F_5, because F_5 involves a fragment on sequence S_1 that is upstream of the corresponding fragment of F_3 on sequence S_1 (Panel (a)). In the graph-representation, such a situation results in no edge assigned to the pair of vertices representing F_3 and F_5. By choosing an appropriate weight function for the edges (see text), the longest path corresponds to the best alignment between the two structures as represented here by thick red arrows.

A path through this graph is a coherent sequence of matching pairs that can be read as an alignment between the two structures. To optimize the structural alignment, we associate a weight to every edge in the graph. An edge (F_1, F_2) connecting two MFPs is assigned a weight $w(F_1, F_2)$ which is given by the sum of two terms:

$$w(F_1, F_2) = w_F(F_1, F_2) + w_C(F_1, F_2)$$

The first term $w_F(F_1, F_2)$ provides a measure of the local similarity of the matching pair F_2. Given the measure of the distance introduced in eq. 1, we can use it to score the similarity between two fragments simply by subtracting it to the value of the m_s threshold

$$S_c(F_2) = m_s - S(i_2, j_2, L_2)$$

This function reaches a maximum if the two fragments are exactly identical ($S(i, j, L) = 0$) and decreases for fragments that are increasingly different. The term $w_F(F_1, F_2)$ is calculated as the score of F_2 ($S_c(F_2)$) multiplied by its length L_2. In case of an overlap between F_1 and F_2, we consider only the length of the non overlapping part of F_2 which is L_1-L_2+i_2-i_1

$$w_F(F_1, F_2) = \begin{cases} L_2 \cdot S_c(F_2) & \text{if } i_2 - i_1 \geq L_1 \\ (L_2 - L_1 + i_2 - i_1) \cdot S_c(F_2) & \text{if } i_2 - i_1 < L_1 \end{cases}.$$

The second term, $w_C(F_1, F_2)$, is a penalization term given by the sum of two contributions: the first penalizing the presence of gaps and the second taking into account the mutual displacement of the two MFPs F_1 and F_2 in the two structures:

$$w_C(F_1, F_2) = G_p \cdot gap\ length + P(D_f(F_1, F_2)).$$

G_P is the gap penalty (set to -0.5 in the current implementation). The term $P(D_f(F_1, F_2))$ penalizes the chaining of

two MFPs that are displaced with respect to one another in the two structures.

For illustrating the function of the $P(D_f(F_1, F_2))$ term, let us consider the case of an alignment including two α-helices. If the two helices have different relative positions in the two structures, the score for their alignment will be penalized by the $P(D_f(F_1, F_2))$ term. The different relative positions can have two different reasons: Either one of the structure undergoes a conformational change moving the two helices with respect to one another (i.e. the alignment is in principle correct) or one of the two helices in one structure is in fact not structurally equivalent to its counterpart in the other structure (i.e. incorrectly aligned). In the first case, both the helices will be part of larger fragments that are structurally equivalent and the penalization introduced by the inclusion of the two helices in the alignment should be balanced by the positive contribution of the MFPs stably surrounding the two helices. If the two helices are not structurally equivalent, then the surrounding MFPs will also not be structurally equivalent thus not giving rise to balancing contributions to the score effectively leading to elimination of the two helices from the alignment.

To achieve the required behaviour of the score, $D_f(F_1, F_2)$ is defined as a measure of the displacement in space of the two matching fragments and is calculated using difference distances between the two fragments. In case the two fragments $F_1 \equiv (i_1, j_1, L_1)$ and $F_2 \equiv (i_1, j_1, L_1)$ have the same length $L = L_1 = L_2$, then D_f is calculated as

$$D_f(F_1, F_2) = \frac{1}{L}\sum_{t=0}^{L-1} \left| d_A(i_1 + t, i_2 + t) - d_B(j_1 + t, j_2 + t) \right|$$

otherwise if L is the minimum between L_1 and L_2 we select in the longest fragment the subfragment of length L yielding to the maximum value of D_f.

P is a truncated linear function calculated as

$$P(D_f) = \begin{cases} 0 & \text{if } D_f < m_{C1} \\ P_C \cdot L_2 \cdot \dfrac{D_f - m_{C1}}{m_{C2} - m_{C1}} & \text{if } m_{C1} \le D_f < m_{C2} \\ P_C \cdot L_2 & \text{if } D_f \ge m_{C2} \end{cases}.$$

In the current implementation the parameters are empirically set to G_P = -0.5, P_C = -5.0, m_{C1} = 1.0, m_{C2} = 4.0. This choice leads to preference for short gaps and longer aligned fragments with fewer hinge regions.

After weights have been assigned to all edges, the best alignment between the two structures can be seen as a '*longest path*' in the weighted graph and is calculated using a dynamic programming algorithm. Since the graph is a DAG the longest path can be calculated in time $O(V+E)$ [47] where V is the number of MFPs and E is the number of edges between them. The number of edges is $O(V^2)$ in the worst case and the number of matching fragments is potentially $O(L^2)$, with L being the average length of the two residue sequences. This means that the worst case complexity of the overall algorithm is $O(L^4)$. Nevertheless, the number of matching fragments is usually much less than L^2 and several heuristics can be used to considerably speed up the algorithm.

An additional issue is taken into account while calculating the best alignment. As discussed above, a strong displacement between two MFPs is identified by a higher value of D_f. This can happen either when the two matching fragments are located on the two sides of a hinge point or if they belong to unrelated and locally similar stretches of residues. The first case can be distinguished from the second by considering that in the case of an hinge point a pair of chained fragments with an high value of D_f will be followed by a sequence of MFPs with lower values. Therefore correct alignments are likely to contain a lower number of chained MFPs with a high value of D_f. Therefore, for each vertex a counter (C_H) for the number of times the D_f term is greater than m_{C2} on the longest path that reaches that vertex, is stored. A maximum threshold for C_H is fixed in the algorithm (M_H) and the algorithm discards paths leading to a value of C_H that is higher than this threshold. In the current implementation, this threshold is fixed to 5. As a result, the alignment provided by the algorithm can cross a hinge point a number of times that must be less than M_H. This heuristic was already used by Ye et al. [18].

Refinement of the alignment

The initial alignment obtained after the chaining of MFPs can be used as a basis for finding additional residue equivalences that can only be detected by checking their consistency with the initial alignment.

At first, for every gap between aligned fragments, the intervening residues are systematically checked to verify if their inclusion is consistent with the rest of the alignment.

For all the aligned fragments, small shifts along the sequence (until the next aligned fragment is reached) are tested in order to correct small offsets in the alignment of periodic structures such as helices that can sometimes occur due to the high local similarity.

Finally, aligned fragments in the vicinity of the N and C-termini are inspected and eventually removed if showing insufficient quality of the alignment.

Technically, all checks are done by evaluating whether or not addition/removal of an equivalent pair of residues improves the scoring function of the genetic algorithm on the error scaled difference distance matrix between the two structures (for details on the scoring function see Schneider [29]).

Adjustable parameters

The only adjustable parameter of the aligner is the *Low limit*. This parameter controls whether or not different distances measured between pairs of equivalent atoms are considered as identical within error. It corresponds to the ε_l parameter in [29] and is set to 2.0 by default. The default value was optimized for the detection of typical domain motions; lower values will enforce a stricter similarity criterion for distances within rigid bodies (higher number of smaller rigid bodies) while larger values will do the opposite (leading to a lower number of rigid bodies with larger size).

Pre-processing step

In order speed up RAPIDO for aligning structures with very similar sequences, a pre-processing step exploiting the fact that sequences can be aligned much more quickly than structures was added to the scheme described above. An initial sequence alignment is in fact performed for all pairs of structures to be aligned. If this sequence alignment reveals a sufficient similarity of the primary sequences (see below), the sequence-based equivalence map is used as a starting point for a preliminary search for rigid bodies. The rigid bodies found are retained and stored as MFPs to be later used by the RAPIDO aligner algorithm. The non-rigid and/or not aligned parts of the two structures are scanned for MFPs using the standard approach described above. The set of MFPs used for the next step of the algorithm (the merging of MFPs) is then created by combining the MFPs from the two sources.

Technically the sequence alignment is carried out using the Smith-Waterman dynamic programming algorithm [48] where a PAM250 [49] matrix is used for amino acids substitutions. If the coverage of the sequence alignment is higher than 90% or both the coverage and identity are higher than 25% the pre-processing step continues with the identification of rigid bodies, otherwise the pre-processing step is aborted and the RAPIDO algorithm is executed with no modifications.

This step is particularly useful in cases like the alignment of structures of GroEL from different organisms, where the time for computation is reduced by 80% using the pre-processing. For the human kinase dataset the pre-processing step is useful in 66% of the alignments (1496 out of 2278) and the computation time is reduced by 70% on the average.

Compilation of the dataset of structures of protein kinase domains

All sequences of human protein kinase domains as defined in the Human Kinome database (http://kinase.com/human/kinome/, [50]) were used to query the database of sequences corresponding to all chains with structures deposited in the Protein Data Bank [51] with the program ssearch34 from the FASTA suite [52]. All hits with E-values less than $1*10^{-90}$ were retained and manually pruned to select only structures with a sequence identity greater than 98%. With this method, for every sequence from the Human Kinome Database, all structures in the PDB that represent the respective protein were identified. For protein kinase sequences with more than one corresponding in the PDB, we then randomly selected one representative structure. The final dataset is composed of 68 structures, resulting in a total of 2278 all-against-all pairwise alignments. The PDB ids of the 68 selected structures and details about the sequence alignments are listed in additional file 3. The version of the Human Kinome database and PDB used in this study were of April 2006.

Authors' contributions

RM conceived and implemented the aligner algorithm, compiled the hinge motions dataset, performed the tests, validated the results and drafted the manuscript. BB compiled the kinase structure dataset and validated the results on that dataset. TRS conceived, designed and coordinated the study and finalized the manuscript. All authors contributed to the discussion of the ideas behind the study. They all read and approved the final manuscript.

Additional material

Additional file 1

Comparison between DALI and RAPIDO on the human kinase structures dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-352-S1.txt]

Additional file 2

results on Fischer's dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-352-S2.doc]

Additional file 3

List of the structures included in the Dataset of human kinase structures.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-352-S3.txt]

Acknowledgements

We would like to thank Dr. Adam Round for the fruitful discussion. This work was supported by grants from Associazione Italiana per la Ricerca sul Cancro (RM, BB, TRS).

References

1. Godzik A: **The structural alignment between two proteins: is there a unique answer?** *Protein Sci* 1996, **5(7):**1325-1338.
2. Lathrop RH: **The protein threading problem with sequence amino acid interaction preferences is NP-complete.** *Protein Eng* 1994, **7(9):**1059-1068.
3. Goldman D, Papadimitriou CH, Istrail S: **Algorithmic Aspects of Protein Structure Similarity.** *focs* 1999:512.
4. Kolodny R, Linial N: **Approximate protein structural alignment in polynomial time.** *Proc Natl Acad Sci USA* 2004, **101(33):**12201-12206.
5. Lemmen C, Lengauer T: **Computational methods for the structural alignment of molecules.** *J Comput Aided Mol Des* 2000, **14(3):**215-232.
6. Sierk ML, Kleywegt GJ: **Deja vu all over again: finding and analyzing protein structure similarities.** *Structure* 2004, **12(12):**2103-2111.
7. Kolodny R, Koehl P, Levitt M: **Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures.** *J Mol Biol* 2005, **346(4):**1173-1188.
8. Taylor WR, Orengo CA: **Protein structure alignment.** *J Mol Biol* 1989, **208(1):**1-22.
9. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **11(9):**739-747.
10. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS: **ProSup: a refined tool for protein structure alignment.** *Protein Eng* 2000, **13(11):**745-752.
11. Szustakowski JD, Weng Z: **Protein structure alignment using a genetic algorithm.** *Proteins* 2000, **38(4):**428-440.
12. Ortiz AR, Strauss CE, Olmea O: **MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.** *Protein Sci* 2002, **11(11):**2606-2621.
13. Ilyin VA, Abyzov A, Leslin CM: **Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point.** *Protein Sci* 2004, **13(7):**1865-1874.
14. Zhang Y, Skolnick J: **TM-align: a protein structure alignment algorithm based on the TM-score.** *Nucleic Acids Res* 2005, **33(7):**2302-2309.
15. Teichert F, Bastolla U, Porto M: **SABERTOOTH: protein structural alignment based on a vectorial structure representation.** *BMC bioinformatics* 2007, **8:**425.
16. Roach J, Sharma S, Kapustina M, Carter CW Jr: **Structure alignment via Delaunay tetrahedralization.** *Proteins* 2005, **60(1):**66-81.
17. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233(1):**123-138.
18. Ye Y, Godzik A: **Flexible structure alignment by chaining aligned fragment pairs allowing twists.** *Bioinformatics* 2003, **19(Suppl 2):**II246-II255.
19. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6(3):**377-385.
20. Krissinel E, Henrick K: **Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.** *Acta Crystallogr D Biol Crystallogr* 2004, **60(Pt 12 Pt 1):**2256-2268.
21. Guda C, Lu S, Scheeff ED, Bourne PE, Shindyalov IN: **CE-MC: a multiple protein structure alignment server.** *Nucleic Acids Res* 2004:W100-103.
22. Lupyan D, Leo-Macias A, Ortiz AR: **A new progressive-iterative algorithm for multiple structure alignment.** *Bioinformatics* 2005, **21(15):**3255-3263.
23. Gerstein M, Lesk AM, Chothia C: **Structural mechanisms for domain movements in proteins.** *Biochemistry* 1994, **33(22):**6739-6749.
24. Gerstein M, Krebs W: **A database of macromolecular motions.** *Nucleic Acids Res* 1998, **26(18):**4280-4290.
25. Xu Z, Horwich AL, Sigler PB: **The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex.** *Nature* 1997, **388(6644):**741-750.
26. Shatsky M, Nussinov R, Wolfson HJ: **Flexible protein alignment and hinge detection.** *Proteins* 2002, **48(2):**242-256.
27. Shatsky M, Nussinov R, Wolfson HJ: **A method for simultaneous alignment of multiple protein structures.** *Proteins* 2004, **56(1):**143-156.
28. Ye Y, Godzik A: **Multiple flexible structure alignment using partial order graphs.** *Bioinformatics* 2005, **21(10):**2362-2369.
29. Schneider TR: **A genetic algorithm for the identification of conformationally invariant regions in protein molecules.** *Acta Crystallogr D Biol Crystallogr* 2002, **58(Pt 2):**195-208.
30. Menke M, Berger B, Cowen L: **Matt: local flexibility aids protein multiple structure alignment.** *PLoS computational biology* 2008, **4(1):**e10.
31. Schneider TR: **Objective comparison of protein structures: error-scaled difference distance matrices.** *Acta Crystallogr D Biol Crystallogr* 2000, **56(Pt 6):**714-721.
32. Schneider TR: **Domain identification by iterative analysis of error-scaled difference distance matrices.** *Acta Crystallogr D Biol Crystallogr* 2004, **60(Pt 12 Pt 1):**2269-2275.
33. Stewart M, Kent HM, McCoy AJ: **The structure of the Q69L mutant of GDP-Ran shows a major conformational change in the switch II loop that accounts for its failure to bind nuclear transport factor 2 (NTF2).** *J Mol Biol* 1998, **284(5):**1517-1527.
34. Vetter IR, Nowak C, Nishimoto T, Kuhlmann J, Wittinghofer A: **Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport.** *Nature* 1999, **398(6722):**39-46.
35. Lee SJ, Matsuura Y, Liu SM, Stewart M: **Structural basis for nuclear import complex dissociation by RanGTP.** *Nature* 2005, **435(7042):**693-696.
36. Nilsson J, Weis K, Kjems J: **The C-terminal extension of the small GTPase Ran is essential for defining the GDP-bound form.** *J Mol Biol* 2002, **318(2):**583-593.
37. Wang J, Boisvert DC: **Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0A resolution.** *J Mol Biol* 2003, **327(4):**843-855.
38. Braig K, Adams PD, Brunger AT: **Conformational variability in the refined structure of the chaperonin GroEL at 2.8 A resolution.** *Nat Struct Biol* 1995, **2(12):**1083-1094.
39. Shimamura T, Koike-Takeshita A, Yokoyama K, Masui R, Murai N, Yoshida M, Taguchi H, Iwata S: **Crystal structure of the native chaperonin complex from Thermus thermophilus revealed unexpected asymmetry at the cis-cavity.** *Structure* 2004, **12(8):**1471-1480.
40. Ranson NA, Farr GW, Roseman AM, Gowen B, Fenton WA, Horwich AL, Saibil HR: **ATP-bound states of GroEL captured by cryo-electron microscopy.** *Cell* 2001, **107(7):**869-879.
41. Branden C, Tooze J: **Introduction to Protein Structure.** Second edition. Garland Publishing, Inc; 1998.
42. Sicheri F, Moarefi I, Kuriyan J: **Crystal structure of the Src family tyrosine kinase Hck.** *Nature* 1997, **385(6617):**602-609.
43. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J: **Structural basis for the autoinhibition of c-Abl tyrosine kinase.** *Cell* 2003, **112(6):**859-871.
44. Xu W, Harrison SC, Eck MJ: **Three-dimensional structure of the tyrosine kinase c-Src.** *Nature* 1997, **385(6617):**595-602.
45. Ogawa A, Takayama Y, Sakai H, Chong KT, Takeuchi S, Nakagawa A, Nada S, Okada M, Tsukihara T: **Structure of the carboxyl-terminal Src kinase, Csk.** *The Journal of biological chemistry* 2002, **277(17):**14351-14354.
46. Fischer D, Elofsson A, Rice D, Eisenberg D: **Assessing the performance of fold recognition methods by means of a comprehensive benchmark.** *Pac Symp Biocomput* 1996:300-318.
47. Cormen TH, Leiserson CE, Rivest RL, Stein C: **Introduction to algorithms.** 2nd edition. Cambridge, Mass.: MIT Press; 2001.
48. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147(1):**195-197.
49. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** *Atlas of Protein Sequence and Structure* 1978, **5(3):**345-352.

50. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2002, **298(5600):**1912-1934.
51. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1):**235-242.
52. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85(8):**2444-2448.