Automatic Determination of Hydrophobic Cores

Mark B Swindells
Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

An algorithm is described for automatically detecting hydrophobic cores in proteins of known structure. Three pieces of information are considered in order to achieve this goal. These are; secondary structure, side chain accessibility and side chain-side chain contacts. Residues are considered to contribute to a core when they adopt a regular secondary structural conformation and have buried side chains which form mainly non-polar contacts with other buried contacts. The efficacy of this method has been assessed by comparing the predictions for interleukin-1 and *Erythrina* trypsin inhibitor structures, with those proposed by different authors on the basis of visual inspection. In these cases the automated procedure shows good agreement with the author definitions despite using only simple descriptives for residue interactions. This method will be useful to all those involved in protein structure analysis by providing an ability to reliably distinguish between buried residues which contribute to a hydrophobic core and those which are only locally buried.

Introduction

One of the most important pieces of structural information that can be derived from a set of atomic coordinates, is the location of helices, strands and loops in the protein. Once these have been defined, our appreciation of the overall topology is significantly enhanced. As a result, numerous groups have developed rapid and objective procedures for automatically assigning regular secondary structure. The most popular of these is without doubt, the DSSP algorithm \(^1\), which is now routinely used in structural analysis. The second piece of information that a researcher may require is a list of the residues which constitute the hydrophobic core. Unfortunately, despite general agreement that most globular proteins possess hydrophobic cores, the definition has remained a steadfastly manual process, with the procedural development receiving only limited attention.

One previously published method uses residue hydrophobicity in combination with side chain-side chain distances to assign an imaginary hydrophobic distance between each pair of residues in the protein \(^2\). Unfortunately the assignments from this algorithm, which is called SCROLL, have significant defects. Firstly the assignment process is dependent on the initial definition of hydrophobicity. As only amino acids which are indisputably hydrophobic are considered, residues such as lysine, which are largely hydrophobic but also contain charged atoms, are automatically neglected. Such consideration is needed for instance in the interleukin-1 (IL-1) topology. Secondly, the hydrophobic cores assigned
seem to be too large. Although the result naturally depends on the subjective definition of a core, it would seem that this method is not effective at distinguishing between clusters which contribute to the core and other false positives. For example, although it is inappropriate to consider adjacent residues in $\beta$ strands as contributing equally to a core, this situation frequently occurs in the SCROIL predictions. Furthermore, there is a surfeit of coil residues in the core. As it is generally accepted that core residues are conserved, between proteins with the same topology, the significant number of coil residues assigned to the core would appear to be inappropriate.

Methods

In this paper, the definition of a hydrophobic core is a set of residues located within regions of regular secondary structure, whose side chains interact with one another and are buried. The assignment of residues to a hydrophobic core involves the preliminary calculation of three pieces of information which are described below:

1) Secondary structure assignments were calculated using the method of Kabsch and Sander (1983). Only residues assigned as helix and strand were retained for further consideration.

2) Side chain-side chain interactions were calculated in the following manner. For each pair of residues $(i,j)$ located within regions of regular secondary structure, atomic contacts were calculated and segregated into hydrophobic contacts $h$ (those between pairs of carbon atoms) and non-hydrophobic contacts $nh$ (all others). A contact was recorded when the distance between two atoms was less than the sum of their van der Waals' radii plus 1Å. Van der Waals' radii were taken from $^3$. Using this information, the total number of hydrophobic ($H$) and non-hydrophobic ($NH$) contacts made by each residue was calculated by summing over all residues. Thus, an interaction between residues $(i,j)$ was recorded when both of the following conditions were met:
   
   \begin{align*}
   a) \quad h_{(i,j)} > nh_{(i,j)} \quad \& \quad h_{(i,j)} > 1 \\
   b) \quad \frac{H_{(i,j)}}{H_{(i)} + NH_{(j)}} > 0.75
   \end{align*}

   These two conditions ensure that the majority of contacts made are hydrophobic, while allowing a certain degree of flexibility.

3) Side chain solvent accessible surface area (ASA) was calculated using an implementation of the $^4$ algorithm written by Simon Hubbard. Atomic van der Waals' radii were taken from $^3$, probe size was 1.4Å and sphere slice was 0.05Å. Relative accessibilities were calculated as a residue's observed ASA relative to its ASA in an Ala-X-Ala dipeptide of extended conformation. Only buried residues were considered to be suitable core residues. As a suitable definition of buried was not known at this time, the cutoff was varied systematically in 1% intervals from 7-15%, in order to explore its effects on the final results.

Although only buried residues could ultimately form the core, it was still of interest to know whether residues which passed these previous criteria were interacting with predominantly accessible or buried residues, as only those forming interactions with a majority of buried residues should be retained for further consideration. Thus, the number of accessible and buried interactions formed by each buried residue was calculated, together with the resultant number of accessible and buried contacts formed by these interactions. Using the selection filter that both the number of buried interactions and contacts must be greater than their accessible counterparts, other residues were removed in an iterative manner until no further change was observed. The remaining residues were classified as belonging to a potential hydrophobic core.
Atomic coordinates were taken from the Protein Data Bank \textsuperscript{5}. Interleukin-1β (PDB code: 1I1B), \textit{Erythrina} trypsin inhibitor (PDB code: 1TIE).

Results

In order to assess the efficacy of this approach, hydrophobic cores were assigned for interleukin-1β (IL-1β) and \textit{Erythrina} trypsin inhibitor (ETI). These proteins were chosen because they adopt the same topology while displaying no detectable sequence similarities. Furthermore, an eighteen residue hydrophobic core had already been proposed for this pair, based on detailed visual analyses\textsuperscript{6}. Thus if the predicted cores for IL-1β and ETI were largely in agreement and these assignments also coincided with the core proposed by Murzin et al.\textsuperscript{6}, it could be concluded that the automated approach was generating realistic assignments.

IL-1β: With a 7% accessibility cutoff, thirteen residues were assigned, eleven of which corresponded with those assigned by the authors. The remaining pair were located at sites which were only two residues distant from other author assigned core residues (Figure 1). This kind of variation was acceptable, as the both approaches were identifying β strand residues which point in the same direction. When the accessibility cutoff was increased to 15%, eighteen residues were assigned. Fourteen of these were in agreement with the author assignments, with the remaining four all occupying locations which were again only two residues distant from the author assigned core.

ETI: When the accessibility was within the range 7-15%, twenty one residues were assigned to the core. In this example all eighteen core residues were detected with three over-predictions. As in IL-1β, over predictions occurred at positions which were only two residues distant from author defined core residues (Figure 1).

Discussion

In this communication, I have described an algorithm which is intended to automatically assign hydrophobic cores. In the past, the success of such algorithms has proved difficult to assess because of the inevitable subjectivity when determining the "correct" answer. To circumvent this problem I have used proteins which adopt similar topologies despite retaining no statistically significant sequence similarities to assess the results. The advantage of this approach is that a similar solution should be recorded, even though the algorithm is being presented with unrelated data (c.f. homologous proteins where the number of conserved residues is significantly higher). As an example I have assigned
hydrophobic cores for IL-1β and ETI. There are practically no residue identities between these proteins and the side chain packing arrangements (assessed by comparing χ₁ and χ₂ angles) are completely different. Furthermore, the accessibilities for structurally equivalent residues vary dramatically in non-core locations. Despite these differences the algorithm identifies a similar pattern of residues for each protein which is extremely encouraging. However, to recommend its general use, the algorithm requires more rigorous testing in order to show that the cutoffs are not biased towards this topology alone. Preliminary results suggest that the algorithm also performs well on globin structures which, unlike IL-1β/ETI, are α helical.

References


<table>
<thead>
<tr>
<th>IL-1β</th>
<th>ETI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5SLNCLRD</td>
<td>13GTYLLP</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>16KSLVMSG</td>
<td>27GGVQLAK</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>23PYELKAL</td>
<td>40PLTVQOS</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>39QVVFMSFV</td>
<td>53GKPIRESR</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IL-1β</th>
<th>ETI</th>
</tr>
</thead>
<tbody>
<tr>
<td>55KIPvALGLK</td>
<td>70DDEVRIIGFA</td>
</tr>
<tr>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>66NLYLSC</td>
<td>86SPWWTV</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>80QLE</td>
<td>101VKLS</td>
</tr>
<tr>
<td>100VFNKIEIN</td>
<td>114FKFEQVS</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Fig 1: Structurally derived sequence alignment for human IL-1β and ETI. The alignment is shown for 88 equivalences. The eighteen author assigned core residues are denoted by + symbols. Residues identified by the automated procedure are denoted by the numbers 7 and 15. These numbers represent the relative accessibility at which they are classified as belonging to the core. In ETI, residue 99 is classified as β strand, but is not superposable with a residue in IL-1β.