Conference Review

Multiple sequence alignments as tools for protein structure and function prediction

Alfonso Valencia*
Protein Design Group, National Centre for Biotechnology, CNB-CSIC, Madrid, Spain

Abstract

Multiple sequence alignments have much to offer to the understanding of protein structure, evolution and function. We are developing approaches to use this information in predicting protein-binding specificity, intra-protein and protein-protein interactions, and in reconstructing protein interaction networks. Copyright © 2003 John Wiley & Sons, Ltd.

Keywords: multiple sequence alignments; protein structure; prediction; protein interactions; correlated mutations; interaction networks

Multiple sequence alignments are a precious repository of the successful evolutionary strategies explored by proteins [24]. We are interested in the development of computational methods able to systematically recover part of this information. A first line of research addressed the detection of positions with patterns of variation characteristic of the internal structure of the corresponding protein families. These ‘tree determinant residues’ are distributed in close proximity to regions dedicated to specific molecular recognition (e.g. binding sites, protein interaction regions) [5]. Indeed, by manipulating these residues it is possible to modulate protein-binding specificity [1] (Figure 1). The possibility of using predicted specificity sites for the prediction of the corresponding molecular functions still remains unexplored. We have initiated a complementary route for the exploration of potential binding regions by training neural networks with proteins with known interaction sites and the corresponding multiple sequence alignments [9].

A second line of research studies the small variations in multiple sequence alignments that may be presented by amino acids that act in association to maintain protein stability against random mutational drift. These correlated positions present a weak correlation with physical proximity in protein structures [11,15] and can be combined with other signals to predict three-dimensional contacts [7,8]. Interestingly, they seem to be more effective in the detection of protein–protein interactions than in the prediction of intra-protein contacts [17] and, as such, can be used to build models of protein complexes [2] (Figure 2).

Sparked by the growing interest in deciphering protein interaction networks (Table 1), we extended both approaches to the prediction of protein interaction partners. The ‘mirror-tree’ method [18] and the ‘in-silico-2-hybrid’ method [19,23] are based on the concepts of tree-determinants and correlated mutations, respectively. Recently, we have systematically evaluated the reliability of the protein interaction network prediction methods, using our previous developments for directly assessing the presence of the interactions in published papers. The predictions compared include those generated by each one of these two methods for the E. coli genome, those by other previous published methods based on information about genome
Swap of function (binding partner) between related proteins. Ras and Ral are very similar proteins of the larger ras super-family that bind to very different effectors: Rlip in the case of Ral, and Ras Binding Domain containing proteins for the ras proteins. The prediction of two residues key for the functional differentiation of the sequences was generated with the SequenceSpace software, based on the tendency of these two positions to contain information specific for each one of the two families [1]. The experimental exchange between the corresponding amino acids in these two positions produced a complete swap of the corresponding substrate-binding specificities. In this case the Ras double mutant bound Rlip and not the ras binding domain containing proteins and the Ral double mutant bound Ras Binding domain proteins and not Rlip [1].

### Table 1. Current genome and sequence-based methods for the prediction of protein–protein interactions

<table>
<thead>
<tr>
<th>Method</th>
<th>References</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene neighbours</td>
<td>3,16</td>
<td>Uses the proximity of genes in bacterial genomes as criteria for the prediction of functional relations</td>
</tr>
<tr>
<td>Gene fusion</td>
<td>6,13,14</td>
<td>Explores the presence of fused genes producing a single peptide chain in some genomes for the prediction of interactions</td>
</tr>
<tr>
<td>Patterns of gene presence</td>
<td>10,20</td>
<td>Genes with a similar distribution in complete genomes are predicted to have related functions</td>
</tr>
<tr>
<td>Domain architecture</td>
<td>12,22</td>
<td>The domain composition of complex proteins is exploited for the prediction of associations between them</td>
</tr>
<tr>
<td>Mirror trees</td>
<td>18</td>
<td>The similarity of the gene trees of different protein families is quantified and used for the prediction of interactions</td>
</tr>
<tr>
<td>In silico-2-hybrid</td>
<td>19</td>
<td>The presence of 'correlated positions' between pairs of positions in pairs of multiple sequence alignments is used as indicative of their potential molecular interaction</td>
</tr>
</tbody>
</table>
Figure 2. Proposed model of polymerization for FtsA. The residues participating in correlations are coloured red and green. The model was selected as the docking model that best fits the distance constraints imposed by the correlated residues. The model is compatible with the position of a peptide able to interrupt dimer formation, which corresponds to an interface region highlighted in yellow in the picture. Reproduced from [2] by permission of John Wiley and Sons Ltd.


Submit your manuscripts at http://www.hindawi.com