Homonyms and synonyms in the Dictionary of Interfaces in Proteins (DIP)

Robert Preißner, Andreas Goede and Cornelius Frömmler

Medical Faculty of the Humboldt University (Charité), Institute of Biochemistry, Monbijoustraße 2A, Berlin D-10117, Germany

Received on January 6, 1999, revised on March 30, 1999, accepted on April 9, 1999

Abstract

Motivation: Should reports on molecular mimicry in particular cases, e.g. responsible for cross-reactivity, be considered as accidental or as a general principle in protein evolution? To answer this question, two types of similarity have to be considered: those in homologues (synonyms) and resemblance between patches from unrelated proteins (homonyms).

Results: All interfaces from known protein structures were collected in a comprehensive databank [Dictionary of Interfaces in Proteins (DIP)]. A fast, sequence-independent, three-dimensional superposition procedure was developed to search automatically for geometrically similar surface areas. Surprisingly, we found a large number of structurally similar interfaces on the surface of unrelated proteins. Even patches from different types of secondary structure were found resembling each other. The putative functional meaning of homonyms is demonstrated with striking examples.

Availability: The downloadable interface files are accessible with a Windows 95 viewer which allows the selection of particular interfaces (http://www.charite.de/ch/biochem/). They can be exported as Brookhaven Protein Data Bank (PDB) files or as subset files for the BOSYM package. The superposition procedure is included in the viewer. Additional interface files are available from the authors on request; data bank screenings can be performed in a collaborative way.

Contact: cornelius.frömmel@charite.de; preissner@rz.hu-berlin.de; andreas.goede@rz.hu-berlin.de

Introduction

The strategy of mimicry for deterrence or camouflage arises at different stages of evolution—think of insects called walking leaves. Man exploits mimicry of transmitters, hormones and substrates in the treatment of diseases by drugs. The occurrence of mimicry at the atomic level at protein surfaces has been demonstrated (Fields et al., 1995). In this case, an anti-idiotypic antibody (E5.2) imitates, by its CDR loops, parts of the molecular surface of lysozyme. About 80% of the atoms of contact areas (~50 atoms) are in similar positions (Fields et al., 1995). Such imitation of antigenic epitopes opens the possibility of vaccination by mimetics.

Antibody–antigen interactions have been subjected to further theoretical studies (Helmer-Citterich et al., 1995) to find correlation to binding of peptides from libraries. Remarkably successful was the design of a cyclic heptapeptide mimicking the CD4-binding loop to MHC and therefore possessing immunosuppressive activity (Satoh et al., 1997). This general strategy towards protein surface mimetics will have to take into account the conserved secondary structures for special functions, as reviewed by Fairlie et al. (1998).

A number of theoretical approaches focus on surface representations useful for analysis of protein–protein interactions (Jones and Thornton, 1997) or docking procedures (Helmer-Citterich and Tramontano, 1994; Meyer et al., 1996). Database systems as a representation of the solvent-accessible surface are efficient for docking procedures (Seid and Kriegl, 1995). The problem of automatic detection of surface motifs is addressed, for example, by the computer vision technique (Fischer et al., 1995), but appears unsolved for the general case in that no three-dimensional structure of the (co-crystallized) complex exists.

For general use, the finding of structurally similar surface motifs had to be possible. This possibility will be checked in this work based on the database of surface patches on secondary structures in proteins.

The similar patches in distantly related proteins (called synonyms here) will be subjected to a detailed analysis because global similarity scores, like an overall r.m.s. deviation, may be inadequate for estimation of homology. This may occur because of the relative shift of domains and/or secondary structures, though local interfaces are conserved.

Accidentally similar patches in unrelated proteins (called homonyms here) may be the basis for binding of similar ligands or simply reflect structural constraints on surfaces opening the view on nature’s modeling kit. In either case, it would be of outstanding interest to discover such homonyms.

Systems and methods

Database

The starting point for this analysis is the Dictionary of Interfaces in Proteins (DIP) (for details, see Preißner et al., 1998).
The content of the database DIP can easily be adapted to the problem concerned (e.g. exclusively members of a family of homologous proteins or proteins binding certain ligands). In this case, the data set was adjusted to avoid redundancy.

**Algorithms**

**Automatic procedure for sequence-independent superposition.** Because the definition of the molecular surface patches (MSPs) exclusively considers spatial atomic neighborhood and chemically incomplete parts have to be considered, a simple algorithm for their superposition was required. It is only sketched here, and further details will be published later. In a first step, the centers of mass of the patches are superimposed, followed by a rotation of one MSP, such that the major directions (least and largest expansions) coincide. This normalization is used in a further step to determine the pairs of atoms between the two patches. Finally, the resulting superposition is expanded for neighboring atoms (Preißner et al., 1998).

**Screening strategy.** The queries for geometrical similarity can be performed using different criteria. For a number of purposes, it may be advisable to focus on exterior patches (i.e. the part of secondary structures towards solvent). According to our experiences with the superposition algorithm, the size range of the considered patches can be restricted to ±5 atoms for medium-sized patches (20–60 atoms). The tolerance in the largest expansion should not exceed 30% of the original size.

**Results and discussion**

The existing retrieval system of the DIP, including an automatic superposition procedure, allows the search for similar MSPs. Here we concentrate on accidentally similar patches in unrelated proteins—called homonyms. Because of their relevance for binding (and lower number), this analysis focused on exterior patches.

**Database screening for homonyms**

Starting with an (exterior) patch consisting of 55 atoms, ~1000 exterior patches of similar size (50–60 atoms) have to be considered (for 300 proteins). This is repeated with 50 arbitrary patches of this size and the results of the 50 000 superpositions are shown in Figure 1. Interestingly, superpositions cluster in two separate distributions (see Figure 1a).
The bimodal distribution is interpreted as randomly matching and on the other hand structurally significant matching with more than two-third of the atoms superimposed at r.m.s. deviations below 1 Å. The distribution of the 1000 best superimposed patches is plotted in detail (see Figure 1b).

**Examples of homonyms**

One surprising result of our search was that we found some structurally analogous interfaces up to 50 atoms in secondary structures of different type. We were able to superimpose up to 24 of 30 atoms from patches of different types of secondary structural elements (SSEs) with an r.m.s. value smaller than 0.5 Å. On average, >50% of these atoms coincide in basic atomic properties like partial charge and hydrophobicity. One curious example should be presented here (Figure 2): one side of an interface between β sheets of immunoglobulin [Brookhaven Protein Data Bank (PDB) code 2FBA] can be superimposed with a part of the surface of the α helix in neutral protease (PDB code 1NPC). Curiously, the backbones show opposite orientation. Further examples of inverse similarity and structural implications are published in Preißner et al. (1997). In this case, for 26 atoms we found an r.m.s. deviation of ~1 Å, although there is absolutely no homology between the sequences.

To demonstrate the improbability of occurrence by chance, we selected a few triples of proteins with resembling patches where ~85% of the atoms are superimposed with r.m.s. values below 0.5 Å (see enlarged marks in Figure 1b, Table 1). Such triples can be constituted by completely different proteins, e.g. ribonuclease F1, phospholipase A2 and maltodextrin binding protein.

**Detection of synonyms**

During detailed analysis of examples, it became visible that our approach detects distant evolutionary relationships between proteins in the database. Three proteins that have completely different functions in the cell are calmodulin, parvalbumin and troponin C, but they all share the EF-hand motif and Interestingly large patches are superimposed very well (for details, see the larger signs in Figure 1b and the data in Table 1). To demonstrate the problem of deciding between homonyms and synonyms, we included a further triple of proteins with calmodulin, another type of parvalbumin and aldose reductase where completely different parts are super-
Table 1

<table>
<thead>
<tr>
<th>protein</th>
<th>PDB-ID</th>
<th>seq.-No. sequence</th>
<th>No. of atoms</th>
<th>rms-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>aldose reductase</td>
<td>1ADS</td>
<td>282-289 SQDMTLLL</td>
<td>54</td>
<td>reference</td>
</tr>
<tr>
<td>parvalbumin</td>
<td>1RTP</td>
<td>8-16 AEDIKKAI</td>
<td>46</td>
<td>0.33</td>
</tr>
<tr>
<td>calmodulin</td>
<td>3CLN</td>
<td>138-145 YEEFVQMM</td>
<td>46</td>
<td>0.49</td>
</tr>
<tr>
<td>glutathion peroxidase</td>
<td>1GP1</td>
<td>185-192 EPDIETLL</td>
<td>53</td>
<td>reference</td>
</tr>
<tr>
<td>manganese superoxide dismutase</td>
<td>1ABM</td>
<td>173-180 RPDLKAI</td>
<td>45</td>
<td>0.44</td>
</tr>
<tr>
<td>maltodextrin binding protein</td>
<td>1DMB</td>
<td>186-200 AGAKGLTFLVDLTK</td>
<td>88</td>
<td>reference</td>
</tr>
<tr>
<td>phospholipase A2</td>
<td>1BP2</td>
<td>40-55 DLRCCQTHDNCYQ</td>
<td>75</td>
<td>0.58</td>
</tr>
<tr>
<td>ribonuclease F1</td>
<td>1FUS</td>
<td>13-28 SQVRAANAACQYYQ</td>
<td>76</td>
<td>0.62</td>
</tr>
<tr>
<td>calmodulin</td>
<td>3CLN</td>
<td>82-92 EEEIREAFVRF</td>
<td>90</td>
<td>reference</td>
</tr>
<tr>
<td>parvalbumin</td>
<td>5PAL</td>
<td>40-50 DAQVKEVFEIL</td>
<td>71</td>
<td>0.57</td>
</tr>
<tr>
<td>troponin C</td>
<td>5TNC</td>
<td>131-141 EEIEDELMKDS</td>
<td>78</td>
<td>0.58</td>
</tr>
<tr>
<td>carboxypeptidase alpha</td>
<td>5CPA</td>
<td>15-28 LDEIYDFMDLLVAQ</td>
<td>88</td>
<td>reference</td>
</tr>
<tr>
<td>arabinose binding protein</td>
<td>1ABE</td>
<td>43-56 GKTIAIDSLAAS</td>
<td>76</td>
<td>0.46</td>
</tr>
<tr>
<td>phosphoglycerate hydrolase</td>
<td>4ENL</td>
<td>221-234 AEAALDILVDAIKA</td>
<td>70</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Fig. 3. Example of similar patches in different proteins responsible for binding of similar nucleotides. Ribbons of phthalate dioxygenase reductase (left; PDB code 2PIA) and ferredoxinNADP reductase (right; PDB code 1FNK) with the superimposed parts of the nucleotides in stick representation. The superimposed patches (31 of 34 atoms; r.m.s. value 0.6 Å) are shown as Connolly surface.

imposed (synonyms; for details, see Table 1). The question of the general functional meaning of similar molecular surface patches cannot be answered at this time, but a striking example for binding of similar cofactors in different proteins is presented. During a database screening, two extended patches from phthalate dioxygenase reductase (PDB
code 2PIA) and ferredoxinNADP reductase (PDB code 1FNR) were superimposed with 31 of 34 atoms (r.m.s. deviation 0.6 Å). These are proteins of different size which share overall sequence identity of ~10%. Phthlate dioxygenase reductase binds FMN, while FAD is bound to ferredoxinNADP reductase. The structural similarity of the superimposed patches is responsible for the nucleotide binding, which occurs in both (Figure 3) and is located in domains that have broadly diverged during evolution (Correll et al., 1993).

Limitations

Limitations of the patch approach result from the fact that binding sites are generally split between few secondary structural elements constituting smaller and larger patches (Peters et al., 1996). There are two ways of dealing with this problem:

1. putting together (small) neighboring patches;
2. limitation on larger patches expressing the individuality of the site.

Therefore, the detection of similarity using the MSPs depends on a minimal size of constituting patches of ~10 atoms at this time.

In the current version of DIP, the definition of the MSPs depends on the accurate definition of the secondary structural elements. Shifts of their borders may lower the matching scores and breaks (e.g. at helix kinks) that can result from small distortions of the backbone sometimes prevent detection.

Summing up, the systematic characterization of structurally similar surface patches may be a promising theoretical approach to ligand binding to proteins because it could result in a thesaurus of antonyms (substrates, cofactors, inhibitors) including a variety of (putative) binding sites. Moreover, recent results on similarity between myoglobin and a receptor protein from Chlamydia which seem to be responsible for autoimmune diseases of the heart (Bachmaier et al., 1999) emphasize the importance of automatic detection of such homonyms at the three-dimensional level.

References