PROTEINS: Structure, Function, and Genetics 52:137–145 (2003)

A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures

Martin Jambon,[1] Anne Imberty,[2] Gilbert Deléage,[1] and Christophe Geourjon[1]*
[1]*Institut de Biologie et Chimie des Protéines (IBCP), Lyon, France*
[2]*Centre Recherches sur les Macromolécules Végétales (CERMAV—Affiliated with Joseph Fourier Université), Grenoble, France*

ABSTRACT **An innovative bioinformatic method has been designed and implemented to detect similar three-dimensional (3D) sites in proteins. This approach allows the comparison of protein structures or substructures and detects local spatial similarities: this method is completely independent from the amino acid sequence and from the backbone structure. In contrast to already existing tools, the basis for this method is a representation of the protein structure by a set of stereochemical groups that are defined independently from the notion of amino acid. An efficient heuristic for finding similarities that uses graphs of triangles of chemical groups to represent the protein structures has been developed. The implementation of this heuristic constitutes a software named SuMo (Surfing the Molecules), which allows the dynamic definition of chemical groups, the selection of sites in the proteins, and the management and screening of databases. To show the relevance of this approach, we focused on two extreme examples illustrating convergent and divergent evolution. In two unrelated serine proteases, SuMo detects one common site, which corresponds to the catalytic triad. In the legume lectins family composed of >100 structures that share similar sequences and folds but may have lost their ability to bind a carbohydrate molecule, SuMo discriminates between functional and non-functional lectins with a selectivity of 96%. The time needed for searching a given site in a protein structure is typically 0.1 s on a PIII 800MHz/Linux computer; thus, in further studies, SuMo will be used to screen the PDB. Proteins 2003;52: 137–145.** © 2003 Wiley-Liss, Inc.

Key words: bioinformatics; 3D structure of proteins; user-defined chemical groups; detection of similar 3D functional sites

INTRODUCTION

Understanding and predicting the function of proteins using bioinformatical tools traditionally uses three levels of knowledge: amino acid sequence, backbone structure, and local arrangement of atoms. Several tools dealing with sequence or main-chain structure are publicly available through the World Wide Web and routinely used by molecular biologists. Blast[1] and Fasta[2] provide efficient ways to extract similar sequences from databases containing millions of sequences. Some other tools help to correlate sequence and function using sequential patterns. The Prosite database[3] consists of human-designed functional signatures that may be searched against a protein sequence. Profile analysis[4] is a technique based on multiple-sequence alignments of homologous sequences and may be used to test whether a sequence belongs to a given family. Pattinprot[5] allows one to search a database for any given pattern, which may have been inferred from multiple-sequence alignments such as those obtained with ClustalW[6] from a set of homologous protein sequences. When a 3D structure of a given protein is available, it is possible to use tools such as the Dali/FSSP server,[7,8] which mainly use the main-chain to find similarities and classify proteins. But all these methods reach their limits because a significant similarity in the sequence or in the backbone structure of two proteins is neither necessary nor sufficient to prove that they share a common biological function.

Inferring biological function from 3D structures of proteins is and will remain a challenging problem, given that it strongly depends on the biological context surrounding every protein molecule in vivo. However, analyzing precisely data provided by crystallographic or NMR experimental studies may show local structural similarities among various proteins, which could be correlated to an already known biological function. Although a lot of efforts have been made in past years to develop surface-matching algorithms,[9] very few methods combine chemical information together with geometry in an efficient manner, and none of them use custom chemical groups as the elementary bricks responsible for biochemical activity. Methodologies based on computer vision heuristics have been developed in the 1990s.[10] These methods are purely geometrical and use discretized representations of molecular surface. Variants and improvements of the original technique select sparse critical points among all points representing the molecular surface[11] and introduce a small number of hinges, allowing flexibility in the docking or matching process.[12] Other tools use the surface representation of the proteins to perform comparisons by other means.[13–15] Chemical environment has been taken successfully into

Grant sponsor : French Ministère de la Recherche.

*Correspondence to: Dr Christophe Geourjon, Laboratoire de Bioinformatique et RMN structurales, Institut de Biologie et Chimie des Protéines 7, Passage du Vercors, 69367 Lyon cedex 07, France.
E-mail: c.geourjon@ibcp.fr

Received 5 July 2002; Accepted 1 November 2002

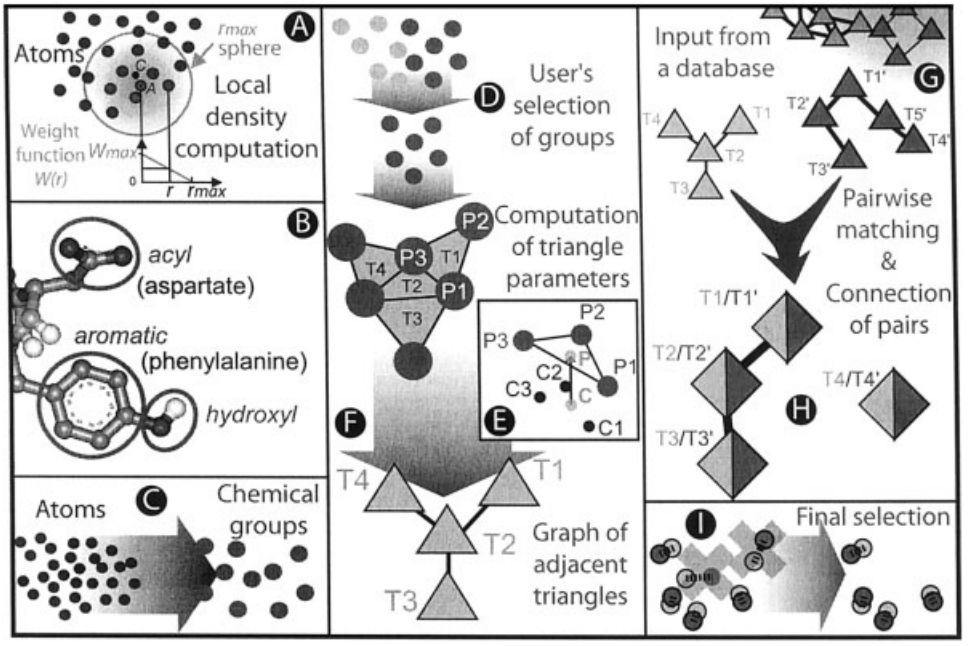

Fig. 1. Major steps in SuMo heuristics. **A:** Local density computation around a given atom A with a plot representing the weight function. **B:** Example of two chemical groups as currently defined. **C:** Reduction of atoms to chemical groups. **D:** Selection of chemical groups according to the user's will. **E:** Computation of parameters associated to a triangle; letters are those used in the text: P_1, P_2, and P_3 are the position of the chemical groups, C_1, C_2, and C_3 are the local centers of mass associated to the chemical groups, *P* and *C* are the centers of these points. *F:* Conversion of the chemical groups to a graph connecting adjacent triangles. *G:* Input of graphs from a database. *H:* Main comparison step. *I:* Refinement step.

account in predictive studies concerning special cases: metal-binding sites[16,17] and sugar-binding sites.[18] Approaches from Russell[19] and Wallace et al.[20] are both based on distance-matching procedures regarding side-chain atoms of some preselected amino acids. The method developed by Wallace et al. is proposed through the PROCAT[21] web server and offers the possibility to match a query structure against a database of enzymatic active site templates. This method is based on a geometric hashing algorithm, in which 3D sites that constitute keys are centered around an atom from a specific amino acid. Our challenge was to provide a generic and extensible tool that returns satisfying results for a large number of protein functions with a minimum knowledge of these functions.

This article describes the original strategy that we have initiated to compare protein structures. This method has been implemented to constitute the SuMo software (acronym for Surfing the Molecules) using the Objective Caml programming language.[22] Our approach allows the comparison of protein full structures or selected sites. The philosophy that we adopted during the development was to consider in a single process all relevant information provided by the protein structure, regarding protein function in its broadest sense. To provide results that validate the biochemical data, it was important to use molecular representations and comparison strategies that fit the intuitive models for chemical interactions; on the other hand, it was also important to choose algorithmic strategies that allow to search the whole PDB for a site in less than 1 day. Finally, the software had to be easy to use and configure.

The capabilities of this tool are illustrated and validated across two extreme kinds of biological problems. First is convergent evolution, with the comparison of two serine proteases having a common biochemical activity but no sequential nor fold similarities. Second, divergent evolution and loss of function due to minor modifications in the protein structure is illustrated by the analysis of the legume lectins family.

METHOD

We have developed a completely new automated approach to the problem of structure-function analysis in proteins. Here, we describe the heuristics that has been designed for performing structural comparison of proteins. To achieve exact reproducibility of the method, all steps were implemented into the SuMo software. The methodology (Fig. 1) is divided into two major steps. First, the PDB file containing the atomic coordinates for a protein structure is converted into a data structure suitable for fast comparison. This representation may be stored into a database dedicated to comparison using SuMo. Then comes the comparison step itself, using preformatted data that may come from this database. Numerical values of parameters have been chosen to obtain best results on average over the studied cases.

Data Preformatting

Before any comparison, the 3D structures of the proteins have to be preformatted (i.e., converted into a representa-

tion that will be used by the comparison heuristics). Because this operation takes usually longer than the comparison itself, the preformatted data may be stored as is into a database. Four successive levels will be considered: (1) atoms, (2) groups of atoms, (3) triangles formed by chemical groups, and (4) vertices in the final graph representing the molecule.

At the atomic level, a parameter D called local atomic density is calculated for each atom A and used later in the comparison process [Fig. 1(A)]. Its purpose is to give a discriminative estimation of the burial of a given atom A:

$$D(A) = \frac{1}{v_{\max}} \sum_{r \le r_{\max}} W(r) m_r \tag{1}$$

where m_r is the mass of the atoms in the sphere of radius r centered on A and $v_{\max}$ is the volume of the sphere of radius $r_{\max}$. A weighting function W is applied to reduce the influence of the peripheral atoms so that D is a continuous function:

$$W(r) = 4 \cdot \left(1 - \frac{r}{r_{\max}}\right) \tag{2}$$

where 4 is the factor that is necessary to obtain a density value that does not depend on $r_{\max}$ in a hypothetical homogeneous medium.

Another parameter C called local center of mass is computed. C is the center of the atoms within the sphere. It is important to notice that vector

$$\overrightarrow{CA} \tag{2a}$$

points toward the exterior of the molecule [Fig. 1(A)].

The next level in the preformatting procedure is the construction of chemical groups [Fig. 1(B)]. The chemical groups are defined for each amino acid as shown in Table I. Some amino acids, such as leucine, are not represented by any group, whereas some others, such as tryptophan, comprise several groups. Every chemical group of the molecule consists of a set of atoms. For a given group, a mean position P of the atoms, a mean position C of the local centers of mass and a mean local density D are computed and recorded. This step reduces the representation of the molecule by a set of groups instead of atoms [Fig. 1(C)].

Chemical groups are then used to build triangles of chemical groups [Fig. 1(E)]. Only triangles with edges shorter than 8 Å are considered. Several parameters are computed for every triangle (P_1, P_2, P_3) . Points C_1, C_2, and C_3 are the local centers of mass of the chemical groups associated to P_1, P_2, and P_3, respectively. P denotes the center of the triangle (P_1, P_2, P_3) and C denotes the center of C_1, C_2, and C_3.The distances between two vertices of a triangle are recorded. The burial of each chemical group is estimated by using the local atomic density. The orientation of the triangle toward the rest of the molecule is estimated by using the scalar triple product of

$$(\overrightarrow{CP}_1, \overrightarrow{CP}_2, \overrightarrow{CP}_3). \tag{2b}$$

TABLE I. Correspondence Between Amino Acids and Chemical Groups As Defined in the Current Input File

Amino acid	Chemical groups (symbolic names)
Alanine	
Arginine	guanidinium
Asparagine	amide
Aspartate	acyl
Cysteine	thiol
Glutamate	acyl
Glutamine	amide
Glycine	glycine
Histidine	aromatic, ammonium
Isoleucine	
Leucine	
Lysine	ammonium
Methionine	thioether
Phenylalanine	aromatic
Proline	proline
Serine	hydroxyl
Threonine	hydroxyl
Tryptophan	aromatic, aromatic, amino
Tyrosine	aromatic, hydroxyl
Valine	

Names used for chemical groups may be freely chosen by the user since SuMo does not associate chemical properties with these names.

The final representation of the molecule [Fig. 1(F)] is obtained by connecting adjacent triangles (i.e., triangles that share exactly two chemical groups) to make a graph in which each triangle forms a vertex.

Comparison of Two Molecules

The comparison of two molecules starts from the graphs of triangles representing the input molecules, possibly coming from a database [Fig. 1(G)]. The comparison heuristic is divided into three steps. First, pairs of similar triangles coming from each of the two molecules are searched and connected according to geometric rules [Fig. 1(H)]. This results in a graph of pairs of similar triangles. The independent subgraphs constitute subsets of pairs of similar triangles that are geometrically consistent. Consistent sets of pairs are called patches. Patches are finally refined [Fig. 1(I)] at the chemical groups level.

The rules for retaining a pair of triangles are the following:

- identity of the chemical groups similar length of the edges ($|\text{length1} - \text{length2}| < 2$ Å) similar depth of each chemical group (local density: $|\text{density1} - \text{density2}| < 0.08$ D/Å^3) similar orientation (scalar triple product: $|\text{x1} - \text{x2}| < 100$ Å^3).

Pairs of similar triangles constitute vertices in a comparison graph [Fig. 1(H)]. To set an edge connecting vertices (T1, T1′) and (T2, T2′), these pairs of triangles have to match the following two conditions:

- Triangle T1 must be adjacent to T2 (in the first molecule), and T1′ must be adjacent to T2′ (in the second

molecule). The angle formed by planes T1 and T2 must be similar to that formed by T1′ and T2′.

Thus, we obtain independent subgraphs that correspond to pairs of similar regions across the molecules. Pairs of triangles are then converted back to pairs of chemical groups. These subsets of pairs of chemical groups are called patches.

To perform this comparison at low cost, the following algorithm has been implemented. First, triangles of each molecule (denoted as Triangles1 and Triangles2) are dispatched into an array (Types1 and Types2) according to the type of the chemical groups that constitute their edges; for example, (acyl, aromatic, hydroxyl) constitutes one type of triangle. This allows one to reduce the cost of the comparison by a factor which grows with k^3 where k is the number of different types of chemical groups. Practically, the cost is reduced by 100 or more. The separation of the different types of triangles is performed as follows:

```
for triangle in Triangles1 do
add triangle to Types1[triangle_type(triangle)]
done
for triangle in Triangles2 do
add triangle to Types2[triangle_type(triangle)]
done
```

Then only triangles of the same type are compared, and pairs formed by similar triangles are added to the list List_of_pairs. The similarity predicate are_similar compares triangles according to the rules given previously. This step uses the following procedure:

```
for type in Triangle_types do
for i in 1 .. |Types1[type]| do
for j in 1 .. |Types2[type]| do
if are_similar (Types1[type][i], Types2[type][j])
then add (Types1[type][i], Types2[type][j]) to List_
of_pairs
done
done
done
```

The construction of the graph of similar pairs of triangles is performed by using the following algorithm, where angle_max has been set to 40°:

```
for (x1, x1′) in List_of_pairs do
for (x2, x2′) in List_of_pairs - {(x1, x1′)} do
if are_adjacent (x1, x2)
and are_adjacent (x1′, x2′)
and |sin (angle (x1, x1′) - angle (x2, x2′))| < sin (an-
gle_max)
then connect ((x1, x1′), (x2, x2′))
done
done
```

The next steps, including search for independent subgraphs, only consider the selected pairs of triangles. The cost of these computations is linear regarding to the number of these pairs and are therefore not described since they are not limitative.

The patches are then refined by using a selection procedure [Fig. 1(I)]. Pairs of chemical groups are superimposed by using root-mean-square deviation (RMSD) minimization. The selection uses the following distance function:

$$dist(g_1, g_2) = \alpha \cdot \|pos(g_1) - pos(g_2)\| + \beta \cdot |D(g_1) - D(g_2)| \quad (3)$$

where $||pos(g_1) - pos(g_2)||$ is the euclidean distance between $g_1$and $g_2$after optimal superposition, and $D(g_1) - D(g_2)$ the difference of local atomic density. α and β are coefficients used to balance the importance of the terms. If $dist(g_1, g_2)$ is higher than a given threshold, then pair (g_1, g_2) is removed from the patch (parameters: $\alpha = 1$; $\beta = 15$; threshold = 2).

Programming Language and Environment

The choice of a well-suited programming language was crucial because the data structures required in SuMo's algorithms are rather complex and numerous. Thus, a language that combines high expressiveness, automatic memory management, and safety together with an efficient and portable compiler was required. The Objective Caml programming language[22–24] was adopted: static type inference, polymorphism, and automatic memory management lead to source code that is several times more concise than the equivalent C/C++ code. The language is essentially based on a functional paradigm, but it also provides mutable data structures such as arrays and records, and a full object-oriented programming system. Therefore, these three programming styles allow one to design efficient algorithms independently from the characteristics of the language. Objective Caml's bytecode and native code compilers and its standard library are available on most common platforms (Unix, Windows, MacOS), making it a language of choice for high-level programming tasks required by modern bioinformatics.

Software Usage and Customization

The software first reads the file containing the definition for chemical groups and then enters an interactive loop. SuMo functions accept optional parameters. Thus, the definition of chemical groups as well as all numeric parameters (cutoff values, thresholds, coefficients, etc.) may be changed if needed.

When working on a specific part of a molecule [Fig. 1(D)], two different selection procedures are available. A restriction consists in totally ignoring a part of the molecule, whereas a true selection only reduces the set of chemical groups that will be used for the final representation of the molecule. A restriction will decrease local atomic densities and a true selection will not. Only true selection has been used in this article.

Performance

For very large molecules, the cost of the preprocessing step grows linearly with the number n of atoms in the structure, whereas the cost of the comparison grows with n^2. However, the comparison of two molecules having 5000 atoms is usually still faster than the preprocessing step

required for these molecules. The reason for this is that only triangles constituted by the same chemical groups are compared.

On a computer with a Pentium III 800MHz processor and the Linux operating system, the current implementation leads to an average time of 10 s for the preprocessing of a 3D structure from the PDB in which strictly redundant sequences have been removed. The comparison of this preprocessed data against a given site (15 chemical groups) is typically 0.1 s. This time has been 0.2 s on average for the comparison of a legume lectin against the selected site. The comparison time of the two proteases (275 vs 237 amino acids) is 2 s.

RESULTS

The results provided by this technique are illustrated by two extreme examples in which the structure-function relationship in these proteins is well known. First, the classical case of convergent evolution of serine proteases illustrates the independence of the method from fold or sequence similarities. The second part illustrates over a larger test set the possible discrimination of functional sites from non-functional sites in the legume lectins family, despite high sequence similarity and low overall RMSD.

Serine Proteases

Subtilisin and γ-chymotrypsin are endoproteases sharing a similar catalytic site: both mechanisms use a catalytic triad formed by an aspartate, a histidine, and a serine. These proteins do not share either a sequence similarity nor a similar fold despite their highly similar active sites. Figure 2(A) shows that the position of these residues has neither the same position nor the same order within the sequence, making it irrelevant to align their sequences. Figures 2(C) and (D) present surface views of both sites and show a striking similarity in the burial of theses residues. Structures 1SBC of subtilisin and 1AFQ of γ-chymotrypsin have been compared by SuMo. The result file is shown in Figure 2(B) and displays one similar region that consists of the catalytic triad (Asp32/Asp102, His64/His57, Ser221/Ser195) and a glycine (Gly127/Gly216), which is also known to play a role in the protease activity.[25] This common patch is ranked first among other patches with lower scores.

Legume Lectins

The structural family of legume lectins is represented by 106 structures publicly available in the PDB (see Ref. 26 for a full review of legume lectin structures). Most of them are functional lectins (i.e., proteins that bind oligosaccharides non-covalently), but some of them have lost the capability to bind sugar at this site despite their overall sequential and structural similarity.[27] Two families of lectin-related proteins without native sugar-binding ability are arcelin and α-amylases inhibitors (four structures), insecticidal proteins that probably lose the carbohydrate-binding activity because of deletions in their genes.[28] In addition, seven structures are available of demetallized lectins (i.e., lectins whose site has been deprived of Ca^{2+} and Zn^{2+}). For example, 1DQ1 and 1DQ2 are two structures of concanavalin A in both native and demetallized forms; although their sequences are identical and their backbone has an RMSD of 0.9 Å for α-carbons, only the first form binds a sugar molecule.

Structure 2PEL of the peanut lectin has been used to represent a functional lectin; its site of interaction with lactose has been selected and compared with every structure within the family. More precisely, all groups that have at least one atom closer than 4 Å to any atom of the ligand were selected. Thus, 10 chemical groups covering 9 amino acids were retained [Fig. 3(B)]. The result of these comparisons is summarized in Figure 3(A). Among all structures, 91 proteins showed at least one similar patch. All of these patches were sugar-binding sites from functional proteins. No patch was detected among the 11 proteins missing the sugar-binding function. Thus, only four functional sites were not detected, and no false positive was obtained. Local conformational changes at the binding site explain the lost of activity in the case of demetallized lectins as shown in Figure 3(C).

DISCUSSION

Validity of the Heuristics

We have developed a method that detects structural similarities in 3D structures of proteins. We have shown that structural similarities are correctly detected in serine proteases that have completely different backbone structures and, therefore, unrelated sequences. Dali[29] finds no similarity between these structures, and it is not possible to propose a valid sequence alignment because of the inversion of catalytic residues in the sequence. In this case, the structural similarity that is automatically detected by SuMo corresponds to a common biochemical function and an identical catalytic mechanism. In practical cases when the structure-function relationship is less understood, we could compare structures of proteins that are known to act as competitors in a biological process: enzymatic catalysis, affinity for a ligand, disruption or activation of biochemical pathways, immunological cross-reactivity, inhibition of cell adhesion, and so forth.

With the example of legume lectins, we showed that SuMo excludes non-functional lectins by comparing them to a functional sugar-binding site despite a high degree of similarity in sequence and in main-chain architecture. This indicates that the position of amino acid side-chains is taken into account by SuMo. It is sensitive enough to detect very subtle conformation changes that are correlated with a loss of function. It can detect proteins that lose some loops involved in carbohydrate binding such as arcelin and α-amylase inhibitors but also the more subtle changes due to demetallization in concanavalin A. In this latter case, the program is sensitive enough to differentiate the structures in which carbohydrate-binding activity was conserved ("locked" or active conformation) from the "unlocked" or inactive ones, the two states differing only by the isomerization of a non-proline peptide bond[30] (difference between 1DQ2 and 1DQ0). On the other hand, SuMo

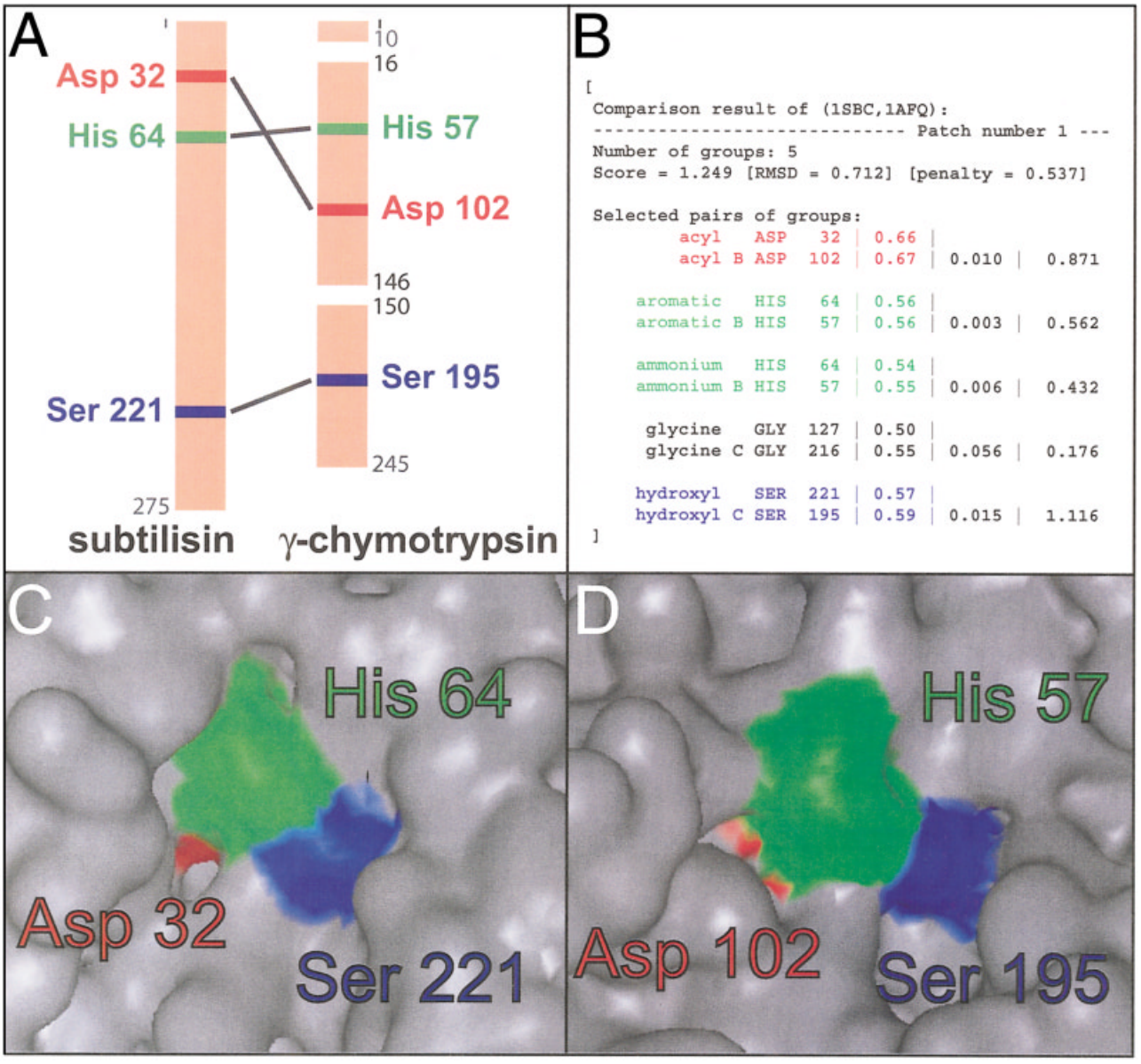

Fig. 2. Comparison of serine proteases: subtilisin, structure 1SBC versus chymotrypsin, structure 1AFQ. **A:** Schematic view of the sequences of both functional forms of the proteins with highlighting of the catalytic triad residues. **B:** Result returned by SuMo. **C:** Surface view of the catalytic triad in 1SBC. **D:** Surface view of the catalytic triad in 1AFQ.

is also flexible enough to ignore minor changes like those depending on the presence or absence of the ligand. It also accepts the changes that are correlated to differences in carbohydrate specificity. Among the hundreds of lectins that have been analyzed and accepted by SuMo criteria, very different specificities are represented (mannose, galactose, complex glycans, etc.). This finding can be explained by the fact that the amino acids in the bottom of the binding site are conserved independently of the carbohydrate (presence of Asp-Asn and aromatic), whereas the molecular basis of the specificity is defined by the loops at the periphery of the binding site.[31]

Only 4% of the functional lectins were not detected: these four structures (1DBN, 1LGB, 1LUL, and 3CNA) have particularities that make it difficult for SuMo to detect the sugar-binding site. In 1DBN, an essential asparagine is replaced by an aspartate residue (Asp 137), but it still binds the sugar molecule by forming a hydrogen bond using one oxygen atom of the carboxyl group. Because the definition that has been used (Table I) for chemical groups does not take into account this feature, the sugar-binding site in this structure has not been detected. The reason of this choice is that hydrogen bond donors and acceptors are very numerous in amino acids and representing each of them by one chemical group leads to bad results. To achieve correct results, chemical groups should be represented by more complex geometrical constructions than a single point in 3D space. That work is

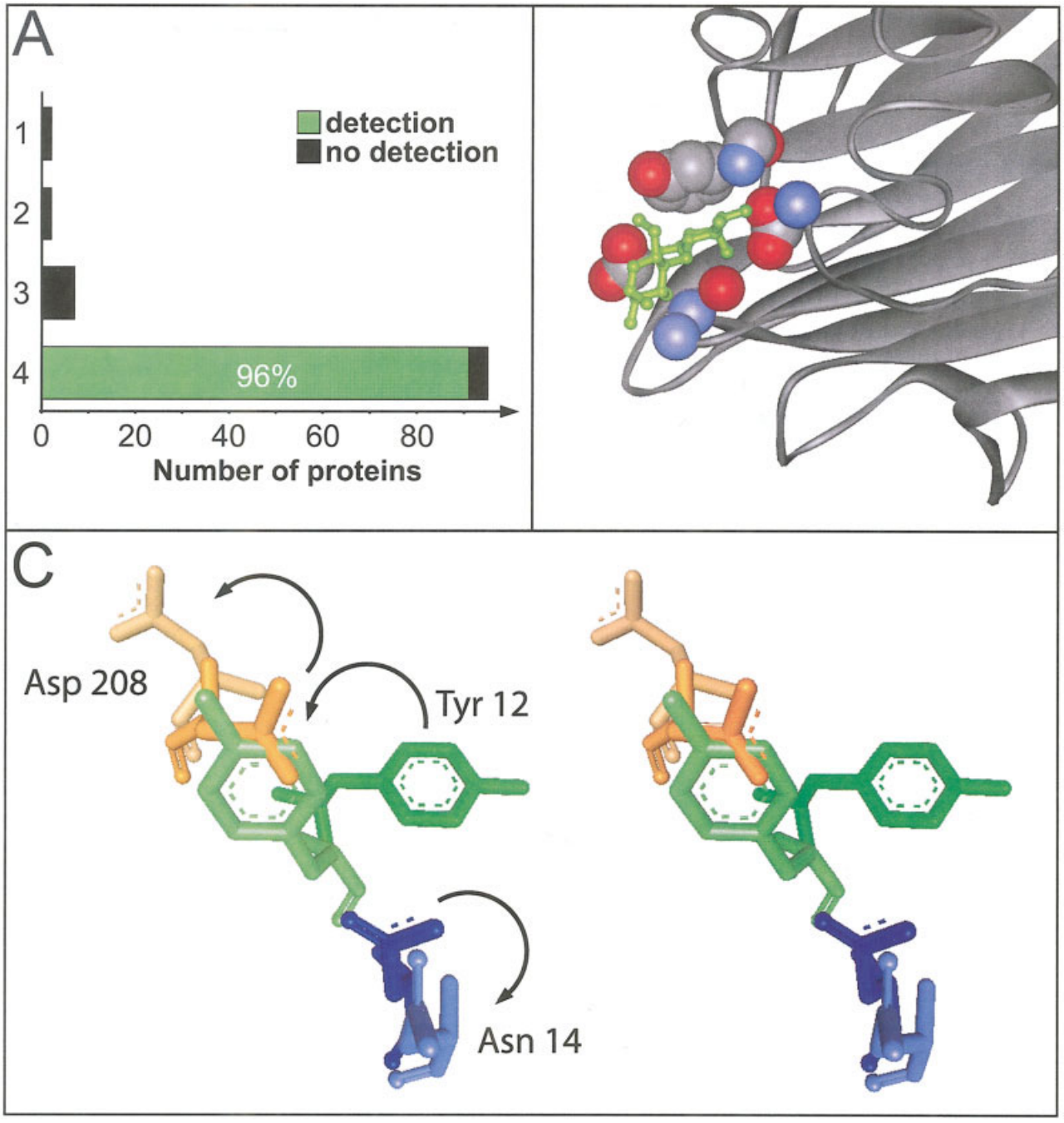

Fig. 3. Screening results of the legume lectin family. **A:** Hits returned by SuMo in the legume lectin family and repartition according to the function of the proteins. (1) arcelin; (2) α-amylases inhibitors; (3) demetallized lectins; (4) functional lectins. Only functional lectins bind an oligosaccharide at the site considered. **B:** View of the chemical groups defining the sugar-binding site in the peanut lectin structure 2PEL. **C:** Stereoview of the amino acids that are essential for sugar binding in concanavalin A before and after demetallization; superposition of α-carbons with 0.9 Å RMSD of 1DQ1 (native form) and 1DQ2 (apoprotein).

currently in progress. The second structure that should have been detected is 1LGB, a lectin that is complexed to another protein near the binding site. By using the default preprocessing of PDB files, the whole complex has been considered; thus, the different local burial as defined by the local density function is higher than in other structures of the family. To check if the negative result is only due to this effect or by a conformational change in the lectin part, the "restrict" option of SuMo has been implemented and used. It offers the possibility to select a part of the structure—the lectin in this case—*before* density computations. By using this modified target, the result was still negative, showing that the binding of the proteic partner leads to a conformational change in the binding

site. Concerning the third problematic structure 1LUL, comments in the PDB file indicate a low resolution for this structure and make us cautious about the interpretation of the data. Interpretation is less clear for the 3CNA structure of concanavalin A. This structure was deposited in the PDB in 1975.

Finally, comparison can be made with other approaches based on statistical analysis of amino acid occurrences in solvent-accessible patches.[18] This approach appeared to be valid for enzymes acting on carbohydrates and for periplasmic carbohydrate-binding proteins, two classes of proteins that generally provide a deep cleft for sugar binding. However, the success for predicting lectin binding site was poor (<30%). The present method yields excellent prediction for protein binding sites as well as for other classes of sugar-binding proteins (in development).

In addition to the examples presented in this article, further positive results have been obtained regarding other biological functions. However, these cases have been studied in less detail. These include sites that bind the following ligands: nucleotides diphosphate (ATP, GTP, etc.), various cations (Ca^{2+}, Zn^{2+}, Fe^{2+}, K^{+}, etc.), Fe_2S_2, and Fe_4S_4 clusters, oligosaccharides.

Methodological and Algorithmic Choices

Russell[19] and Wallace et al.[20] have reported methods that handle similar problems. All approaches, including ours, use 3D matching of labeled points representing functional elements. These elements are either amino acid side-chains, critical atoms, or arbitrary clusters of atoms. Russell's method and ours allow the identification of similar regions in a pair of protein structures without prior knowledge of the functional elements.

The use of chemical groups to represent elementary bricks instead of amino acids to understand molecular functions is essential but not possible if only the sequence of the protein is known. Therefore, the primary structure of proteins is usually modeled by a sequence of amino acids. However, amino acids are composed of several critical groups that may or may not be important, depending on the structural context (Table I). The representation of a macromolecule in Russell's approach[19] allows at most one functional group per amino acid side-chain. The TESS algorithm from Wallace et al.[20] is based on the identification of equivalent atoms, without clustering them into larger groups. However, knowing the 3D structure of proteins allows us to model proteins with chemical groups, covalent bonds, and other interactions independently from the concept of amino acid and even from the concept of individual atom. The definition for chemical groups was chosen to provide the most relevant results in the studied cases but may not be perfect because it depends on the heuristics: introducing a large number of poorly located chemical information leads to bad results and poor performance. This is the reason why hydrophobic chains are not yet taken into account in the current definition of SuMo chemical groups, except in the case of aromatic rings. However, we do not restrict the amino acids to the most conserved ones as in Russell's approach.[19] This point is essential if the protein has no homologue or if optional secondary functional sites exist in a given family of proteins. The full independence from the notion of amino acids allows several extensions of the method. For instance, backbone hydrogen bond donors and acceptors may be taken into account similarly to some side-chain chemical groups; even structures of molecules other than proteins could be considered without major change in the heuristic and could be compared with any other kind of molecule with defined structure.

The choice of using triplets of chemical groups as the basic information for the comparison has been made for several reasons: (a) a triangle may be associated to a number of parameters that ensures that a given triangle contains an amount of information that makes it much more rare than a single group represented by a point. It stands for a minimal representation of a local environment, including an oriented plan; (b) a specific biological function is rarely fulfilled by only one or two chemical groups; (c) the basic information that consists of the chemical group type and its position is kept along all comparison steps. Further work may be performed to consider the chemical groups as solid objects with a given symmetry; (d) adjacent triangles are easy to cluster to represent larger regions of molecules; (e) the all-discrete approach for mapping the spatial properties of the molecules allows the use of efficient heuristics on graphs.

A limit to the representation of a molecule by a single graph of 3D located objects (such as points or triangles) is the difficulty to mix well-located and numerous objects (such as hydrogen bonds) with less located but sparser objects (such as clusters of three positively charged amino acids).

As opposed to most heuristics in the field of structural bioinformatics, the burial of atoms is not estimated by using accessible surface area (ASA) calculation, but a notion of local atomic density. In analogy to immerged bodies, the ASA would correspond to the emerged part of a floating body and be null for any object under the surface, whereas the density calculation may be seen as a measure of the depth of any object, even non-floating ones. Figures 2(C) and (D) show that aspartate in the catalytic triad of serine proteases is almost completely buried, suggesting that crucial residues may be essential for protein function, and this even if they lie below the surface of the molecule. This kind of depth estimation is also essential for providing a vector that is roughly orthogonal to the molecular surface; these vectors are used to estimate the angle formed between a given triplet of chemical groups and the surface.

Because the method may be used to perform a large number of comparisons, especially when one of the compared elements is a small site, several database-based strategies may be used in near future. In a drug design process, screening the PDB for a given 3D site should return a list of potential cross-reactants and help to design target sites for a potential drug. The opposite should also be possible: given a database of ligand-binding sites, one could screen it with a full protein structure and predict the

location of ligand-binding sites. This kind of screening could be used in a structural genomics approach when more and more structures with poor functional data are available.

ACKNOWLEDGMENTS

This work was supported by grants to M.J. from the French Ministère de la Recherche. The authors thank Dr. Roland Montserret for English improvement. Thanks are due to the Caml development team at INRIA Rocquencourt for their high-quality work.

REFERENCES

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
2. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990;183:63–98.
3. Bairoch A. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 1991;Suppl 19:2241–2245.
4. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84: 4355–4358.
5. Blanchet C. Logiciel MPSA et ressources bioinformatiques client-serveur Web dédiés à l'analyse de séquences de protéine. Lyon, France: Université Claude Bernard, Lyon 1; 1999.
6. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
7. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
8. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
9. Via A, Ferre F, Brannetti B, Helmer-Citterich M. Protein surface similarities: a survey of methods to describe and compare protein surfaces. Cell Mol Life Sci 2000;57:1970–1977.
10. Fischer D, Bachar O, Nussinov R, Wolfson H. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. J Biomol Struct Dyn 1992;9:769–789.
11. Lin SL, Nussinov R, Fischer D, Wolfson HJ. Molecular surface representations by sparse critical points. Proteins 1994;18:94–101.
12. Sandak B, Nussinov R, Wolfson HJ. An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. Comput Appl Biosci 1995;11:87–99.
13. Preissner R, Goede A, Frommel C. Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. J Mol Biol 1998;280:535–550.
14. Preissner R, Goede A, Frommel C. Homonyms and synonyms in the dictionary of interfaces in proteins (DIP). Bioinformatics 1999;15:832–836.
15. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
16. Karlin S, Zhu ZY. Characterizations of diverse residue clusters in protein three-dimensional structures. Proc Natl Acad Sci USA 1996;93:8344–8349.
17. Wei L, Altman RB. Recognizing protein binding sites using statistical descriptions of their 3D environments. Pac Symp Biocomput 1998:497–508.
18. Taroni C, Jones S, Thornton JM. Analysis and prediction of carbohydrate binding sites. Protein Eng 2000;13:89–98.
19. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
20. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
21. Wallace AC, Thornton JM. PROCAT, a database of 3D enzyme active site templates http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html.
22. Leroy X, Doligez D, Garrigue J, Rémy D, Vouillon J. The Objective Caml system. Software and documentation on the web, http://caml.inria.fr/ocaml/. 1996.
23. Weis P, Leroy X. Le langage Caml: Dunod; 1999.
24. Chailloux E, Manoury P, Pagano B. Développement d'applications avec Objective Caml: O'Reilly; 2000.
25. Voet D, Voet J. Biochemistry. New York: Wiley and Sons; 1997. p 389–400.
26. Loris R, Hamelryck T, Bouckaert J, Wyns L. Legume lectin structure. Biochim Biophys Acta 1998:9–36.
27. Lis H, Sharon N. Lectins: carbohydrate-specific proteins that mediate cellular recognition. Chem Rev 1998;98:637–674.
28. Mirkov TE, Wahlstrom JM, Hagiwara K, Finardi-Filho F, Kjemtrup S, Chrispeels MJ. Evolutionary relationships among proteins in the phytohemagglutinin-arcelin-α-inhibitor family of the common bean and its relatives. Plant Mol Biol 1994;26:1103–1113.
29. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
30. Bouckaert J, Dewallef Y, Poortmans F, Wyns L, Loris R. The structural features of concanavalin A governing non-proline peptide isomerization. J Biol Chem 2000;275:19778–19787.
31. Young NM, Oomen RP. Analysis of sequence variation among legume lectins. A ring of hypervariable residues forms the perimeter of the carbohydrate binding site. J Mol Biol 1992;228:924–934.