An 0(n2) algorithm for 3D substructure matching of proteins

Xavier Pennec, Nicholas Ayache

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An $O(n^2)$ Algorithm for 3D Substructure Matching of Proteins

Xavier PENNEC, Nicholas AYACHE

N° 2274

Mai 1994

PROGRAMME 4

Robotique,

image

et vision

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

An $O(n^2)$ Algorithm for 3D Substructure Matching of Proteins

Xavier PENNEC, Nicholas AYACHE

Programme 4 — Robotique, image et vision
Projet Epidaure

Rapport de recherche n° 2274 — Mai 1994 — 17 pages

Abstract: Most biological actions of proteins depend on some typical parts of their three-dimensional structure, called 3D *motifs*. To automatically discover corresponding 3D motifs between proteins, we propose a new 3D substructure matching algorithm based on geometric hashing techniques. The key feature of the method is the introduction of a *3D reference frame* attached to each amino acid. This allows to compute for every couple of amino acids 6 invariants, and therefore drastically reduce the complexity of both the preprocessing and recognition stages of geometric hashing, typically to $O(n^2)$.

A thorough analysis of the propagation of the errors in preprocessing and the introduction of extended Kalman filtering insure efficiency and robustness during the recognition stage.

Our experimental results confirm the validity of the approach, and the remarkable stability of the 6 rigid invariants used for matching. We believe that this new algorithm, because of the reduction of the algorithmic complexity it implies, will allow the systematic comparison of very large structures in the near future.

Key-words: Protein Structural Matching, 3D Protein Modeling, 3D Rigid Matching, Registration, Geometric Hashing.

(Résumé : tsvp)

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

Un Algorithme en $O(n^2)$ pour le Recalage de sous-Structures dans les Proteines

Résumé : La plupart des actions biologiques des proteines dépendent de parties précises de leur structure 3D que l'on nomme *motifs*. Pour decouvrir automatiquement les motifs 3D se correspondant entre proteines, nous proposons un nouvel algorithme de reconnaissance et recalage de sous-structures 3D basé sur des techniques de geometric hashing. Le point clé de la méthode est l'introduction d'une *base Euclidienne 3D* attachée à chaque acide aminé, ce qui permet de calculer 6 invariants par couple d'acides aminés, et par là même de réduire drastiquement la complexité, typiquement en $O(n^2)$.

Une analyse précise de la propagation des erreurs dans l'étape de prétraitement et l'introduction du filtre de Kalman étendu assurent l'efficacité et la robustesse lors de la reconnaissance.

Nos résultats expérimentaux confirment la validité de l'approche et la stabilité remarquable des 6 invariants utilisés pour le matching. Nous croyons que ce nouvel algorithme permettra dans un futur proche la comparaison systématique de très grandes structures grâce à la réduction de la complexité qu'il autorise.

Mots-clé : Recalage Structurel de Proteines, Modelisation 3D de Proteines, Recalage Rigide 3D, Reconnaissance, Geometric Hashing.

1 Introduction

1.1 Background

Most biological actions of proteins, such as catalysis or regulation of the genetic message (transcription, maturation, etc...) depend on some typical parts of their three-dimensional structure, called 3D *structural* or *binding motifs*.

Proteins with similar 3D motifs often show similar biological properties, and it is therefore highly desirable to search for similar 3D motifs between proteins [BT91]. Since proteins are composed of thousands of atoms, this search requires efficient and fully automated methods. Applications in biology and medicine are immense, ranging from drug design and disease prediction to protein design and engineering, without forgetting research on the understanding of action mechanisms.

The proteins structure is typically modeled at three scales:

- the primary structure is the linear succession of amino acids,
- the secondary structure is the *local* organization of the chain into structural motifs,
- and the tertiary structure is the *complete* description of the positions of atoms in space, and can reveal the presence of 3D non linear motifs.

Most of existent techniques for comparing proteins deal with the comparison of the primary structure only, using character string comparison algorithms, and especially dynamic programming based ones (see [Mye91, LC91]). These approaches can hardly find most of the structural or binding motifs, whose nature is intrinsically 3D. For instance, polypeptide chains that form a specific structural motif frequently show no or very low sequence homology, even when they are associated with the same specific function [BT91, page 99]. Hence, only the comparison of the tertiary structures of the chains can reveal 3D correspondences.

Moreover, the availability through international data banks of an everyday increasing number of 3D structures allows for the systematic search of motifs present in large sets proteins, provided that efficient and fast 3D substructure matching algorithms do exist.

In this spirit, Fischer and his colleagues [FBNW92, FNW92] have exploited the geometric hashing paradigm previously introduced in computer vision by Lamdan and Wolfson [LW88, Wol90]. They proposed substructure matching methods based on preprocessing and recognition algorithms of complexity $O(n^3)$, where n is the number of atoms of interest (either in the motif or in the protein). A key point of their approach is the possibility to refer to 2 rigid invariants (the "distance coordinates") of any atom of the protein with respect to two other atoms picked arbitrarily as forming a geometric "basis". The results reported in their publications were encouraging, and motivated our work.

1.2 An $O(n^2)$ 3D substructure matching algorithm

Following their pioneering work, our main idea was to reduce the size of a "basis" from two to a single atom. To achieve this goal, we introduce a *3D reference frame* attached to each amino acid. Doing this, we can now choose a single amino acid as a basis, and compute 6 rigid invariants (the parameters of translation and rotation) attached to any other amino acid. This allows to drastically reduce the complexity of both the preprocessing and recognition stages of geometric hashing, typically from $O(n^3)$ to $O(n^2)$.

A thorough analysis of the propagation of the errors in geometric hashing due to one of the authors of this paper [Pen93] and the introduction of extended Kalman filtering for the clustering of found transformations [Aya91, GA92] guided our implementation to insure efficiency and robustness of the approach.

Our experimental results confirm the validity of the approach, and the remarkable stability of the 6 rigid invariants used for matching. We believe that this new algorithm, because of the reduction of the algorithmic complexity it implies, will allow the systematic comparison of very large structures in the near future.

Our paper is organized as follows: first we detail the reference frame attached to each amino-acid, and then the new geometric hashing algorithm we propose for matching. Third we report our experimental study, and finally we present a few potential extensions of our work.

2 Protein structure modeling

Proteins are composed of possibly several chains of amino acids linked each others by peptide bonds. Three groups of atoms in each amino acid constitute the backbone of the chain : the central atom C_α to which are attached on each side an NH group and a carbonyl group $C' = O$ (see figure 1). The residue R, also bound to the C_α, characterizes the nature of the amino acid but does not take part to the backbone of the chain.

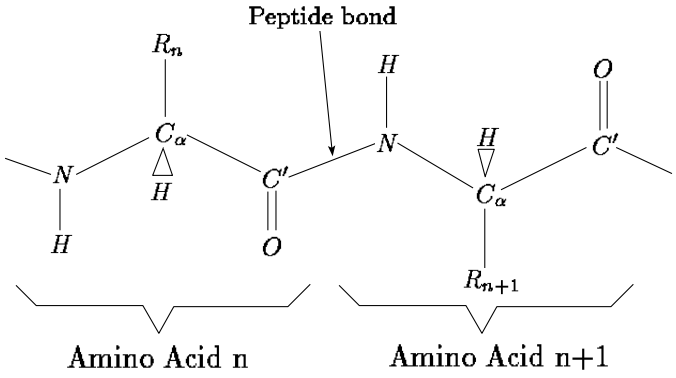
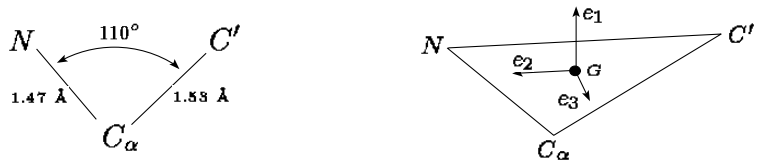


Figure 1: Structure of a Protein chain : a peptide bond links amino acids between each others in a linear way.

Topologically, the backbone of the chain is linear, but its geometry is very complex. Even if rotations are allowed around the bonds $C_\alpha - C'$ and $C_\alpha - N$, and hence the geometry of the chain is weakly constrained, the geometry of the atoms attached to the C_α is perfectly determined. In particular, the three atoms N, C_α, C' form a known triangle from which we can define a basis (see figure 2). We shall see later the function of this basis. It is sufficient to say for the moment that this basis uniquely defines the position and orientation of the amino acid in space. We will hence model an amino acid by a couple (point, trihedron) with possibly a label (the type of amino acid), and a protein by the set of these couples.

The structure comparison problem is thus stated as follows : given two such sets, find all rigid transformations that match a minimum number of amino



$$G = \frac{1}{3}(N + C' + C_\alpha) \qquad e_1 = \frac{C'N \times C'C_\alpha}{||C'N \times C'C_\alpha||} \qquad e_2 = \frac{C'N}{||C'N||} \qquad e_3 = e_1 \times e_2$$

Figure 2: Geometry of an amino acid around the C_α and definition of a basis.

acids of the two structures. The problem can be extended to the comparison of a target molecule with a data-base of proteins. We will see that the geometric hashing paradigm is especially well suited for such a research.

3 Matching Proteins

The problem we are confronted with is very close to recognition problems in volume image analysis, especially in the medical field. In this case, one has to process points extracted from surfaces with their associated Frenet trihedron (see [Thi93, Aya93, GA91]). So in both cases, the model adopted to reduce the data is a set of couples (point, trihedron). Classical techniques rely on model-based approach of object recognition (for a survey, see [Gri92]). Given a data-base of modeled objects (called models), the aim is to recognize in a scene what objects are present, and how they are placed. The simplest problem where the data-base is reduced to only one object is called registration.

3.1 The Geometric Hashing algorithm

The geometric hashing algorithm was introduced by Lamdan and Wolfson for model based recognition in computer vision [LW88, Wol90]. The basic idea is to store in a data-base at preprocessing time a redundant representation of models, based on local features to allow for occlusion, and invariant by rigid transformation. Doing so, the representation of the scene computed at recognition time will present some similarities with that of some objects of

the data-base. Accumulating these evidences will allow the recognition and registration of objects present in the scene and in the data-base.

- *Invariant description*

  Since there is equivalence between finding a rigid transformation and finding corresponding model and scene frames, the main idea is to represent an object by the coordinates of every points in every possible reference frame. In our case, we do have a point and a trihedron associated with each amino acid that uniquely defines a reference frame (see figure 2). Hence, given an amino acid, the coordinates of other amino acid in its reference frame are invariant by rigid transformation. In order to use the informations relative to the orientation of the second amino acid, we add to the 3 coordinates the 3 angular values between the edges of the 2 triangles ($N - C_\alpha - C'$). For each couple of amino acid, we obtain a 6-dimensional vector invariant by rigid transformation. The representation is then the set of every couple of amino acid of the protein, each one being an entry for the hash table, with the 6D invariant vector as index.

- *Preprocessing*

  In order to optimize the access to the representation for recognition time, the geometric hashing algorithm uses a hash table for storing models. Indeed, given one object, just compute the 6D invariant vector associated with each possible couple (reference frame, amino acid), and set it as an index in a 6D hash table for the couple. Each model is processed independently, but stored in the same hash table. The complexity of the step is $O(Mm^2)$, where M is the number of models and m the mean number of amino acids par model. The complexity in space for the hash table is the same since it only depends on the number of entries. This step is performed without any knowledge of the scene to be matched and hence can be done once for all.

- *Recognition*

  As in the preprocessing stage, choose an amino acid as the reference frame. For every other amino acid of the scene, compute the 6D invariant vector and retrieve the compatible model couples (reference frame, amino

acid) in quasi constant time thanks to the hash table. During the process, maintain a list of the model reference frames found, and for each one accumulate the number of compatible couples found. This will be the score. The simple matching of a model and scene reference frames is sufficient for computing a rigid transformation, but every compatible couple brings up some additional information that can be used to refine it using an extended Kalman filter.

The process is repeated for each amino acid considered as the scene reference frame. The output is the list of model and scene matching reference frames with their associated score and rigid transformation. We only keep the matches with a score above a threshold. This parameter is either static or dynamically adapted during the algorithm. It is also possible the keep a fixed number of matches (usually the best ones).

Considering that the access to the hash table is done in constant time, which is true for well distributed tables, the complexity of the whole stage is $O(n^2)$ where n is the number of amino acids in the scene.

- *Error handling*

  Due to the resolution of the X-ray determination of protein structure, conformational deformations and even structural differences between molecules that induce different constraints on the motif, one has to deal with errors in atom positions. A previous study on the propagation of errors within geometric hashing shows that, considering a bounded error ε on the position of every atom, one can determine bounds on the error for the invariant vector used to hash [Pen93]. We do only provide here the practical bounds we used and redirect the interested reader to the original paper for further developments. Let r be radius of the inscribed circle to the triangle used to form the reference frame, and d be the inter-distance of the two amino acids within the considered couple. The error in the coordinates of the amino acid in the reference frame (first part of the 6D vector) is bounded by $\varepsilon_{gh} = \varepsilon \left(1 + \frac{d}{3r}\right)$. A strong supplementary constraint is given by the bound $\varepsilon_d = 2\,\varepsilon$ on the distance d. Concerning the 3 angular values (second part of the 6D vector), a theoretical error bound on each angle is $\varepsilon_\theta = \frac{2\,\varepsilon}{l} + O(\varepsilon^2)$ where l is the length of the

considered triangle edge, but half this bound is a good practical value. As numerical values, we compute that $r = 0.38$ Å and use an average value $l = 1.7$ Å for triangle edges. Each bucket of the hash table intersecting the volume error of a 6D invariant vector is defined as an entry for the couple. Since the hash table can be coarsely discetized, we verify at recognition time that found couples do have compatible invariant vectors.

3.2 Clustering and extension

From now on, we use a probabilistic scheme. Hence, each amino acid frame is given an associated covariance matrix, which is propagated through the computations. The original covariance matrix is diagonal with a value σ_d^2 for positioning and σ_θ^2 for angular error. In our implementation we use $\sigma_d = 1$ Å and $\sigma_\theta = 0.3$ rad.

- *Clustering*

  We have to aggregate the redundant informations obtained from the geometric hashing stage. In order not to mix up the different possible models, we split the list of matching reference frames into one list for each model. For each such list, we consider the first match as a cluster, and examine every other match as follows:

  - Decide with a χ^2 test on Malahanobis distance between the two transformations if they are compatible.
  - If so, merge them together using the extended Kalman filter : the cluster transformation is updated with the new 3 pairs of atoms matched together.

  Repeat the process until every match is incorporated into a cluster. The complexity is $O(k_m k_c)$ with a number k_m of matches and k_c of clusters on output, but k_c is quasi constant in practice.

- *Extension*

  Clusters must now be checked and their matching list extended. This is done using an alignment test: using the rigid transformation previously

determined, the model is mapped onto the scene and the possible matches verified. For the sake of simplicity, only the position of the C_α is now considered in each amino acid. Each C_α of the model is examined as follows.

- Map the model C_α onto the scene and search for the closest C_α of the scene. In order to keep the algorithm symmetrical between the model and the scene, map back the scene C_α to the model and verify that the original model C_α is its closest neighbor. If not, reject the model C_α.
- Compute the Malahanobis distance between the transformed model C_α and the scene one, and decide using a χ^2 test if this match is valid. If not, reject the model C_α.
- Update the rigid transformation of the cluster with this new match using the extended Kalman filter.

Seeking the closest neighbor is performed using k-D trees [PS86]. The complexity of constructing a k-D tree is $O(n \log n)$ with $O(n)$ storage. The search for a closest neighbor is sub-linear, and almost constant in practice. Hence, the whole stage has a complexity $O(n \log n + nk)$.

3.3 Algorithm analysis

- *Simplifications* : The following heuristics can help speeding up the algorithm :
  - Only keep couples with inter-distances under a threshold (typically around 20 Å). The underlying justification is that amino acids of the motif are usually close in space.
  - Label the reference frames (or even every amino acid) with their residue name. Indeed, since we only seek in this step some initial matches, we do not need to obtain every match and it can be sufficient to consider amino acids with the same residue. This is however not used in our experiments.

- *Complexity* : Let m be the mean number of amino acids in the models, M the number of models, n the number of amino acids in the scene, and k the mean number of clusters found by model. The theoretical complexity is $O(Mm^2)$ for the preprocessing stage (comprising the k-D trees preprocessing of models). The whole recognition stage is in $O(n^2 + Mkn)$. In fact, as far as the number of models M is low compared to n, the real complexity is dominated by the geometric hashing stage ($O(n^2)$).

- *Parameters* : There is a small number of parameters that need to be adjusted in the algorithm, such as the bound ε on atom coordinates errors and the variances σ_d and σ_θ on the associated frame, the threshold on the score for geometric hashing, and the thresholds χ_{Clust} and χ_{Verif} used for χ^2 tests. We evaluate the values of these parameters in a learning step: knowing two matched motifs, we register them and compute the variances and the bound ε. Using these informations, we compute in a second step the minimal thresholds. In order to keep some control over the algorithm, we choose to parameterize the variances by the bound ε in a linear way and keep the χ values constant. Hence, in most cases, the only parameters we have to play with are ε and the minimal score for geometric hashing.

4 Experimental results

For all our experiments, we use the atoms coordinates of proteins provided by Brookhaven National Laboratory's Protein Data Bank [BKWM77, ABBK87]. Visualization is done using the RasMol program of R. Sayle [SB92]. For rigid transformations, we provide the translation and the rotation vector (see [Aya91]). Experiments were done with and without the distance constraint without any difference. The labeling scheme was not used, and hence amino acids are not discriminated in the process.

4.1 Detection of a structural motif : the Helix-Turn-Helix motif

Structural motifs can be defined as the super-secondary structure. They are the simple combination of a few secondary structure elements. Some of them are associated with particular functions or are simple parts of larger structural and functional assemblies. Therefore, the Helix-Turn-Helix motif is responsible for the binding of DNA within many procaryotic proteins. Some of them bind tightly to the DNA at a promoter of a gene, preventing RNA polymerase from fixing and hence blocking the initiation of the transcription. They are repressors. Conversely, activators bind next to the promoter and help polymerase to bind.

We choose to compare the tryptophan repressor for E. Coli (PDB code 2WRP [LZSO88]) and phage 434 CRO (PDB code 2CRO [MWH89]), whose Helix-Turn-Helix sequence are known to be [BM89, HA90] :

Protein	*Position*	*Sequence*		
$2CRO$	$15-37$	MT	$QTELATKAGV$	$KQQSIQLIEAG$
$2WRP$	$66-88$	MS	$QRELKNELGA$	$GIATITRGSNS$

In this experiment, we compare the entire proteins in order to find the motif. The two corresponding sequences are correctly matched, and a few other non linear matches are found (table 1).

```
Cluster 1 : score 27
Rotation Vector : -2.53546   0.291995  0.345864
Translation     : -17.9746  -1.90902  -0.211789

   ---------------       15 MET - 66 MET       24 ALA - 75 LEU       33 LEU - 84 ARG
     2CRO - 2WRP         16 THR - 67 SER       25 GLY - 76 GLY       34 ILE - 85 GLY
   ---------------       17 GLN - 68 GLN       26 VAL - 77 ALA       35 GLU - 86 SER
    9 ARG - 63 ARG       18 THR - 69 ARG       27 LYS - 78 GLY       36 ALA - 87 ASN
      ...      ...       19 GLU - 70 GLU       28 GLN - 79 ILE       37 GLY - 88 SER
   11 ILE - 64 GLY       20 LEU - 71 LEU       29 GLN - 80 ALA          ...      ...
      ...      ...       21 ALA - 72 LYS       30 SER - 81 THR       43 ARG - 50 ALA
   15 MET - 66 MET       22 THR - 73 ASN       31 ILE - 82 ILE          ...      ...
   16 THR - 67 SER       23 LYS - 74 GLU       32 GLN - 83 THR       45 LEU - 53 THR
                                                                     ---------------
```

Table 1: Detected matches between the proteins 2CRO and 2WRP. The output of the algorithm is compacted for a better understanding.

We do not assess any biological significance to these 4 supplementary matches. They are only the result of a geometric matching. On the other hand, the fact that other matches do not score more than 21 shows that the HTH motif is the main common structural motif between the two proteins. Images of the proteins and of the registration are provided in figure 3, 4 and 5.

4.2 Detection of a binding site : the heme pocket

The globin family collect proteins of many different organisms. Their amino acid homology can be as low as 16%. However, their 3D structure is still related and constitute the globin fold. It is mainly composed of α helices that form a pocket for the active site which in myoglobins and hemoglobins binds a heme group. The motif constituted by the amino acids binding the heme in 9 globins was extensively studied by Lesk *et al.* [LC80].

We define the motif by 15 of the 19 non sequential positions of amino acids that make contact with the heme in the α subunit of human hemoglobin (PDB code 4HHB, chain α [FPSF84]). We choose the 15 positions given by Lesk *et al.* as being present in 7 or more globins. We search for it within the β subunit of the same protein (chain β), horse hemoglobin (PDB code 2DHB [Pa], chain α and β), myoglobin (PDB code 4MBN [Tak84]), and sea lamprey cyanohemoglobin (PDB code 2LHB [HHL85]). The resulting matches are given in table 2. We present in figure 6, 7 and 8 images of the motif, the β chain of human hemoglobin, and the registration of the two structures.

The matching with horse hemoglobin (2DHB) was done with the two chains together and the two matches were perfectly identified. For the sea lamprey cyanohemoglobin (2LHB), the coordinates deposed by [HHL85] are not the same as those used by to Lesk *et al.*: a SER is deleted after 96-SER, and 98-LEU, 99-ARG were inserted (see remark 6 of the PDB entry). Hence residue 101 from PDB correspond to residue 100 of Lesk *et al.* and so on.

In these experiments, no other match scoring more than 5 was found. This tends to show that the correct recognition greatly emerges from the noise of false positives, and hence our scheme is very robust. This means in particular that the modeling of amino acids is justified.

4HHB α (motif)	4HHB β	2DHB α	2DHB β	4MBN	2LHB
42 TYR	41 PHE	42 TYR	41 PHE	42 LYS	51 PHE
43 PHE	42 PHE	43 PHE	42 PHE	43 PHE	52 PHE
58 HIS	63 HIS	58 HIS	63 HIS	64 HIS	73 HIS
61 LYS	66 LYS	61 LYS	66 LYS	67 THR	76 ARG
62 VAL	67 VAL	62 VAL	67 VAL	68 VAL	77 ILE
65 ALA	70 ALA	65 GLY	70 SER	71 ALA	80 ALA
66 LEU	71 PHE	66 LEU	71 PHE	72 LEU	81 VAL
83 LEU	88 LEU	83 LEU	88 LEU	89 LEU	101 LEU
86 LEU	91 LEU	86 LEU	91 LEU	92 SER	104 LYS
87 HIS	92 HIS	87 HIS	92 HIS	93 HIS	105 HIS
91 LEU	96 LEU	91 LEU	96 LEU	97 HIS	109 PHE
93 VAL	98 VAL	93 VAL	98 VAL	99 ILE	111 VAL
97 ASN	102 ASN	97 ASN	102 ASN	103 TYR	115 TYR
98 PHE	103 PHE	98 PHE	103 PHE	104 LEU	116 PHE
101 LEU	106 LEU	101 LEU	106 LEU	107 ILE	119 LEU
T_x	0.142694	0.493322	-0.04313	3.44391	10.1853
T_y	-1.59334	1.36633	-0.39684	14.5187	18.8873
T_z	2.60057	0.162643	2.35737	-6.03178	-14.3859
R_x	-3.14764	0.004991	3.11788	1.13209	1.73283
R_y	-0.007530	0.033864	-0.03635	-0.672907	1.58384
R_z	0.0802843	-0.016262	-0.127903	0.015026	0.113322

Table 2: Matches and transformations resulting from the algorithm. The transformations map the motif onto the scanned structures.

5 Conclusion

Modeling amino acids by the 3 atoms of their backbone allows to define a complete and unique associated reference frame. Every couple of amino acids has hence 6 invariants for rigid transformations that we use in a geometric hashing scheme to discover initial matches. These are clustered, verified and extended. The error inherent to the problem is integrated in the process, thanks to an error analysis and Extended Kalman Filter. Experiments confirm the validity, efficiency and robustness of our approach.

Future work will be articulated upon three axes. We plan to automatize the adjustment of the algorithm parameters based on a statistical study of the invariants. A second direction would be the use of a probabilistic scheme for

geometric hashing (see [RH93]). Last but not least, we hope to demonstrate the possibility of automatically discovering and extracting the model of an unknown motif common to a given group of proteins.

References

[ABBK87] E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, and J. Weng. Protein data bank. In F.H. Allen, G. Bergerhoff, and R. Sievers, editors, *Crystallographic Databases - Information Contents, Software Systems, Scientific Applications*, pages 107–132, Data Commission of the Int. Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[Aya91] N. Ayache. *Artificial Vision for Mobile robots - Stereo-vision and Multisensor Perception.* MIT-Press, 1991.

[Aya93] N. Ayache. Computer vision applied to 3d medical images: Results, trends and future challenges. In *Int. Symp. on Robotic Research, Hidden Valley, Pensylvania, USA*, 1993. Also as INRIA Research Report No 2050.

[BKWM77] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. of Mol. Bio.*, 112:535–542, 1977.

[BM89] R.G. Brennan and B.W. Matthews. The helix-turn-helix dna binding motif. *J. Biol. Chem.*, 264:286–290, 1989.

[BT91] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing, 1991.

[FBNW92] D. Fischer, O. Bachar, R. Nussinov, and H Wolfson. An efficient automated computer vision based technique for detection of three dimensionnal structural motifs in proteins. *J. of Biomolecular Structures and Dynamics*, 9(4):769–789, 1992.

[FNW92] O. Fischer, R. Nussinov, and H Wolfson. 3d substructure matching in protein molecules. In *Combinatorial Pattern matching 92 - Lect. Notes in Comp. Sci. no 644*, pages 136–150, Spinger Verlag, 1992.

[FPSF84] G. Fermi, M.F. Perutz, B. Shaanan, and R. Fourme. The crystal structure of human deoxyhaemoglobon at 1.74 angstroms resolution. *J. Mol. Bio.*, 175:159, 1984.

[GA91] A. Guéziec and N. Ayache. *Smoothing and Matching of 3D Space Curves.* Research Report 1544, INRIA, 1991.

[GA92] A. Guéziec and N. Ayache. Smoothing and matching of 3d space curves. In *Proceedings of the Second European Conference on Computer Vision*, Santa Margherita Ligure, Italy, 1992.

[Gri92] W.E.L. Grimson. *Object Recognition by Computer - The role of Geometric Constraints.* MIT Press, 1992.

[HA90] S.C. Harrison and A.K. Aggarwal. Dna recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.*, 59:933–969, 1990.

[HHL85] R.B. Honzatko, W.A. Hendrickson, and W.E. Love. Refinement of a molecular model for lamprey hemoglobin from petromyzon marinus. *J. Mol. Bio.*, 184:147, 1985.

[LC80] A.M. Lesk and C. Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Bio.*, 136:225–270, 1980.

[LC91] B Lacroix and J.J. Codani. *Techniques Informatiques pour la Cartographie Physique du Génome Humain.* Research Report 1560, INRIA, 1991.

[LW88] Y. Lamdan and H.J. Wolfson. Geometric hashing : A general and efficiant model-based recognition scheme. In *Proc. of Second ICCV*, pages 238–289, 1988.

[LZSO88] C.L. Lawson, R.G. Zhang, R.W. Schevitz, Z. Otwinowski, A. Joachimiak, and P.B. Siegler. Flexibility of the dna-binding domains of trp repressor. *Proteins Struct., Funct., Genet.*, 3:18, 1988.

[MWH89] A. Mondragon, C. Wolberger, and S.C. Harrison. Structure of phage 434 cro protein at 2.35 angstroms resolution. *J. of Mol. Bio.*, 205:179, 1989.

[Mye91] E.W. Myers. *An overview of sequence comparison algorithms in molecular biology.* Technical Report TR 91-29, Univ. of Arizona, Dep. of Comp. Sci., 1991.

[Pa] M.F. Perutz and al. Private communication.

[Pen93] X. Pennec. *Correctness and Robustness of 3D Rigid Matching with Bounded Sensor Error.* Research report 2111, INRIA, 1993.

[PS86] F. Preparata and M. Shamos. *Computational Geometry, An Introduction.* Spinger Verlag, 1986.

[RH93] I. Rigoutsos and R. Hummel. Distributed bayesian object recognition. In *Proceedings of Int. Conf on Comput. Vis. and Pat. Recog*, pages 180–186, IEEE Computer Society Press, 1993.

[SB92] R. Sayle and A. Bissel. Rasmol: A program for fast realistic rendering of molecular structures with shadows. In *Proceedings of the 10th Eurographics UK'92 Conference*, Universtity of Edinburg,Scotland, 1992.

[Tak84] T. Takano. Refinement of myoglobin and cytochrome c. In S.R. Hall and T. Ashsida, editors, *Methods ans Applications in Crystallographic Computing*, page 262, Oxford University Press, Oxford, England, 1984.

[Thi93] J.P. Thirion. *New Feature Points based on geometric invariants for 3D Image Registration.* Research Report 1901, INRIA, 1993. To appear in ECCV'94.

[Wol90] H.J. Wolfson. Model-based recognition by geometric hashing. In O. Faugeras, editor, *Proc. of 1st Europ. Conf. on Comput. Vision (ECCV 90)*, pages 526–536, Springer-Verlag, 1990. Lecture Note in Computer Science 427.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399

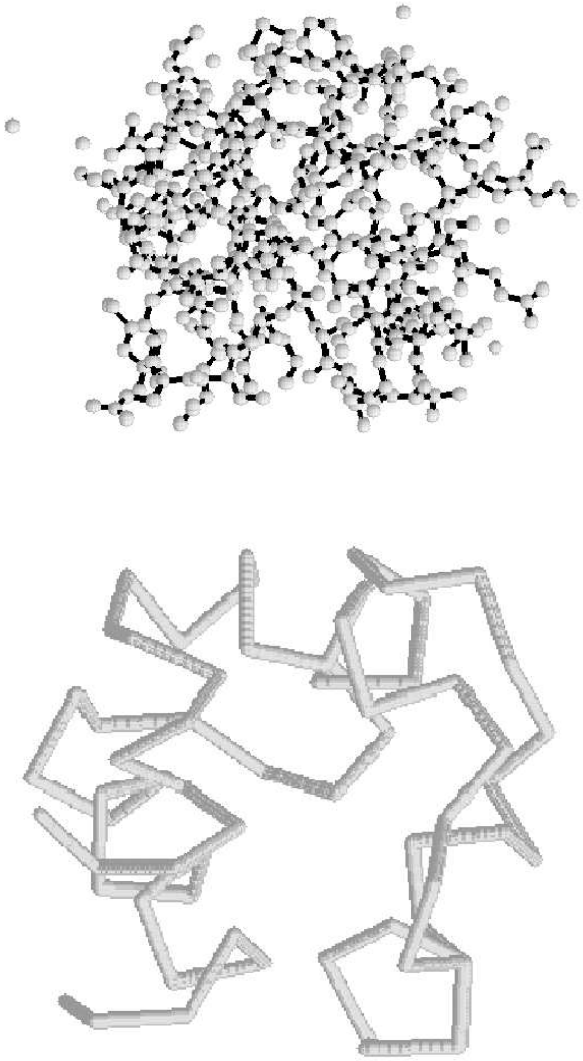


Figure 3: The Cro protein of phage 434 displayed in balls and sticks (top) and its backbone (bottom)

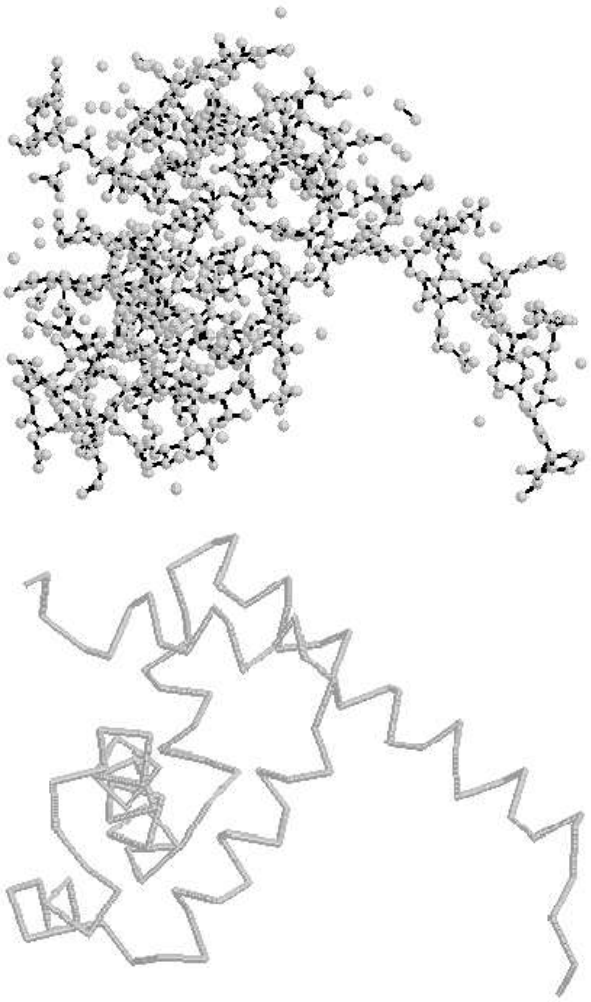


Figure 4: The tryptophan repressor of E. Coli displayed in balls and sticks (top) and its backbone (bottom)

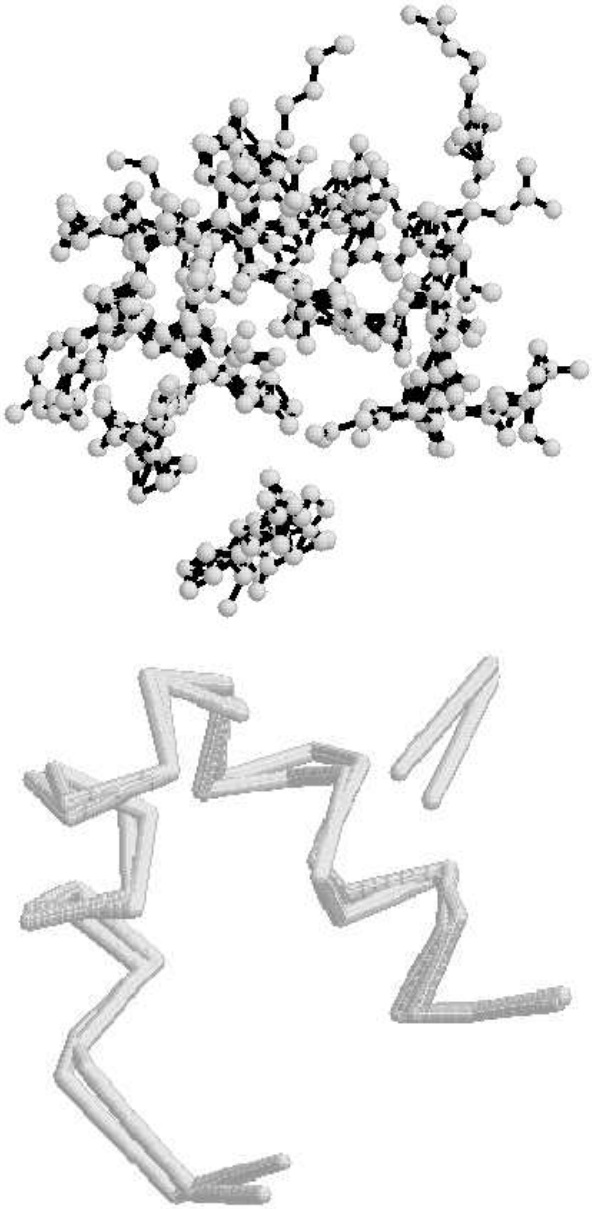


Figure 5: Registered proteins 2CRO and 2WRP. For the sake of visibility, we only show the matched residues. Balls and sticks (top) and backbone (bottom)

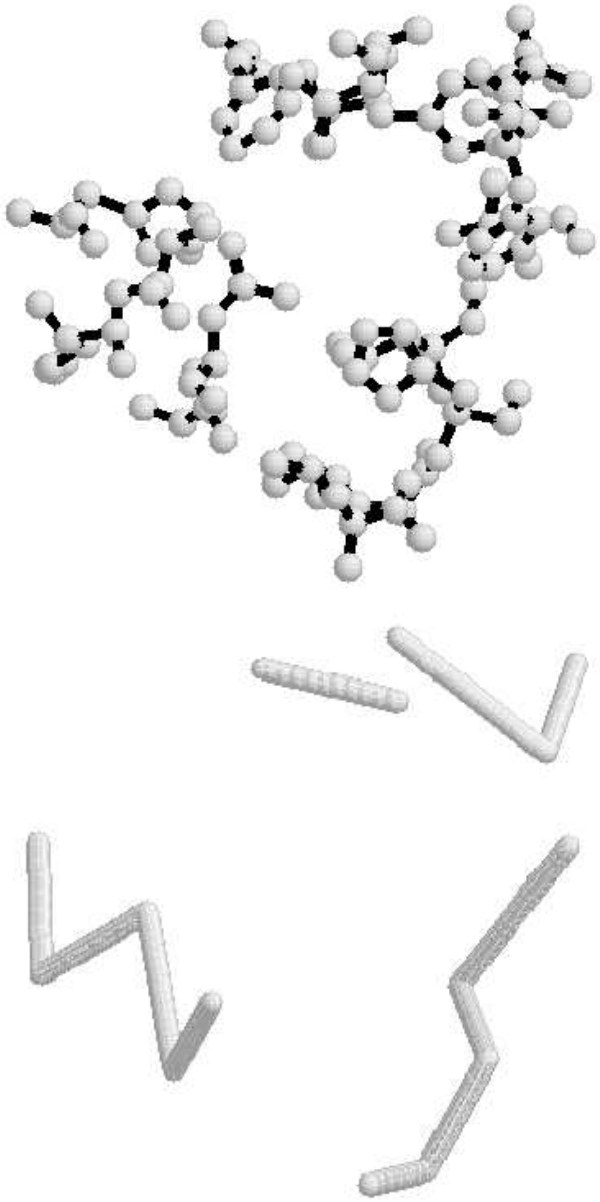


Figure 6: The motif extracted from the α chain of human hemoglobin (4HHB) displayed in balls and sticks (left) and its backbone (right)

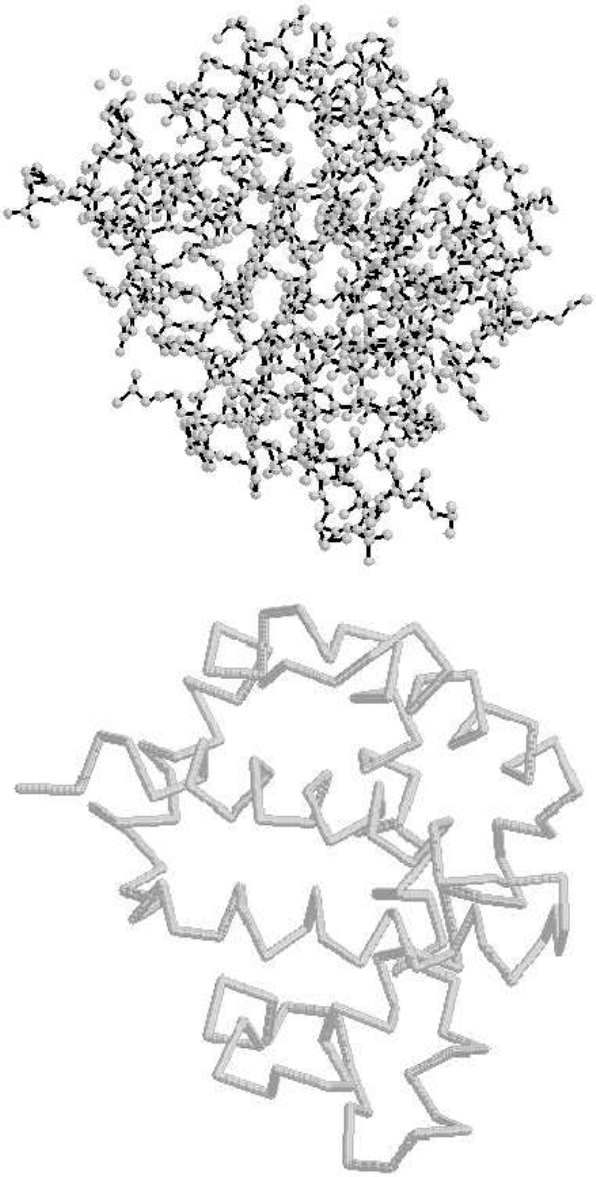


Figure 7: The β chain of human hemoglobin (4HHB) displayed in balls and sticks (left) and its backbone (right)

Figure 8: Motif and chain 4HHB β registered. For the sake of visibility, we only show the matched residues. Balls and sticks (left) and backbone (right)