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

Xavier PENNEC, Nicholas AYACHE
INRIA, B.P. 93, 2004 route des Lucioles
06902 SOPHIA ANTIPOLIS Cedex, FRANCE
Email: Xavier.Pennec@sophia.inria.fr, Nicholas.Ayache@sophia.inria.fr
Fax: (33) 93 65 76 69

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 ensure 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.

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 protein structure is typically modeled at three different 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 [Mye91, LC91]. These approaches can not easily 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 $\mathrm{O}(n^3)$, 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 $\mathrm{O}(n^2)$ 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 [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 describe 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.

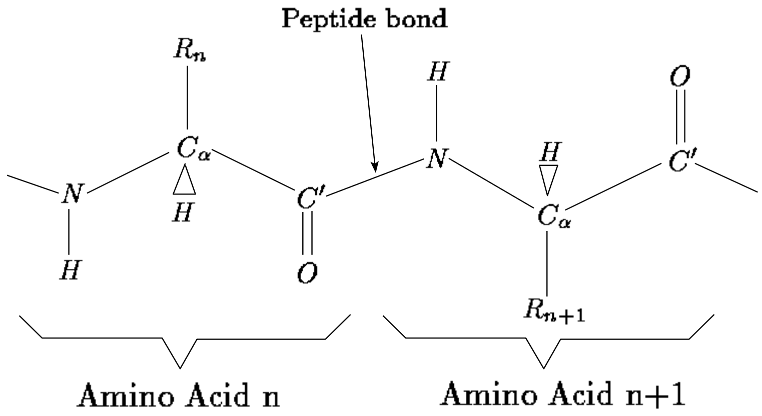
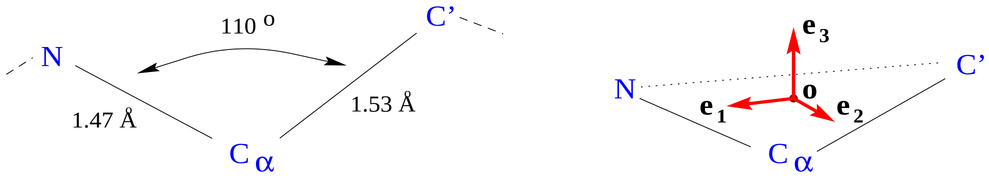


Figure 1: Structure of a Protein chain : a peptide bond links amino acids between each others in a linear way.

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



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

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

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 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 [Thi93, Aya93, GA93]. 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. See for instance [Gri90] for a survey. 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 simply matching or sometimes 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*
  In our case, local features are frames. However, in the absence of labels, any model frame can be matched with any scene frame. To obtain an invariant description, we thus have to consider binary constraints between frames. Indeed, a couple of frames has 6 invariants given by the parameters of the rigid transformation from the last frame to the first (figure 3). For ease of work, we use in fact the 3 coordinates of the translation (expressed in frame i), and the 3 angular values between the corresponding axis of the two frames. These values will constitute the 6D invariant vector of a frame couple. In order to deal with occlusion, the representation of one frame has to be redundant: each frame will then be associated with any other frame of the object to compute the set of 6D invariant vectors characterizing this reference frame (see for example figure 4). The global representation is then the set of every frame couple of the model, 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

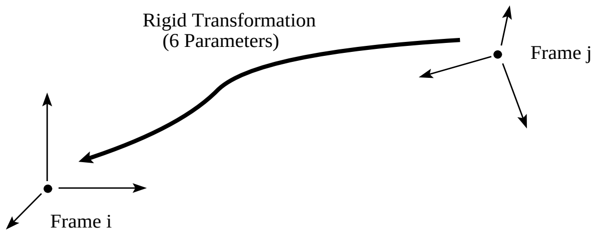


Figure 3: If frame i is considered as the basis of the Euclidean space (reference frame), the 6 parameters of the rigid transformation mapping frame j on the origin (frame i) are invariants.

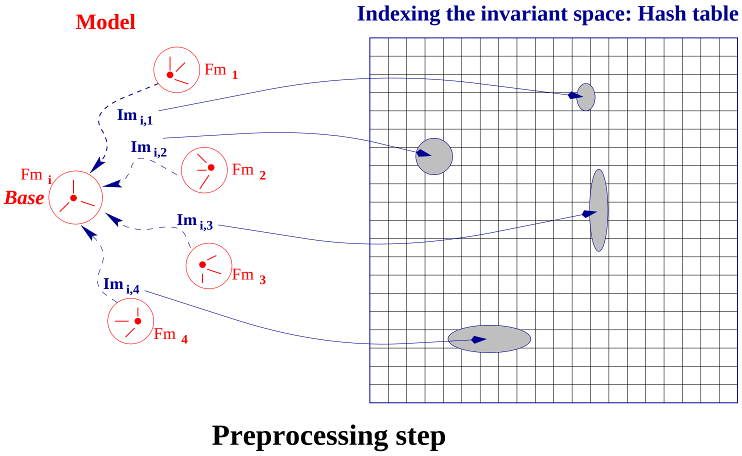


Figure 4: Preprocessing: the 6D invariant vector associated with every couple of frames is computed with its error zone (see Error handling section) and used as an index for the couple in the hash table. For the sake of simplicity, we only deal here with the representation of model frame i (Fm_i).

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, other model frame), 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 $O(M.m^2)$, where M is the number of models and m the mean number of amino acids per model. Typical values for m range from 15 for motifs to a few hundreds for big proteins. 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*
  Choose a reference frame; for each different scene frame, compute the 6D invariant vector and retrieve the compatible model couples (reference frame, other model frame) in quasi constant time thanks to the hash table. During the process, maintain a

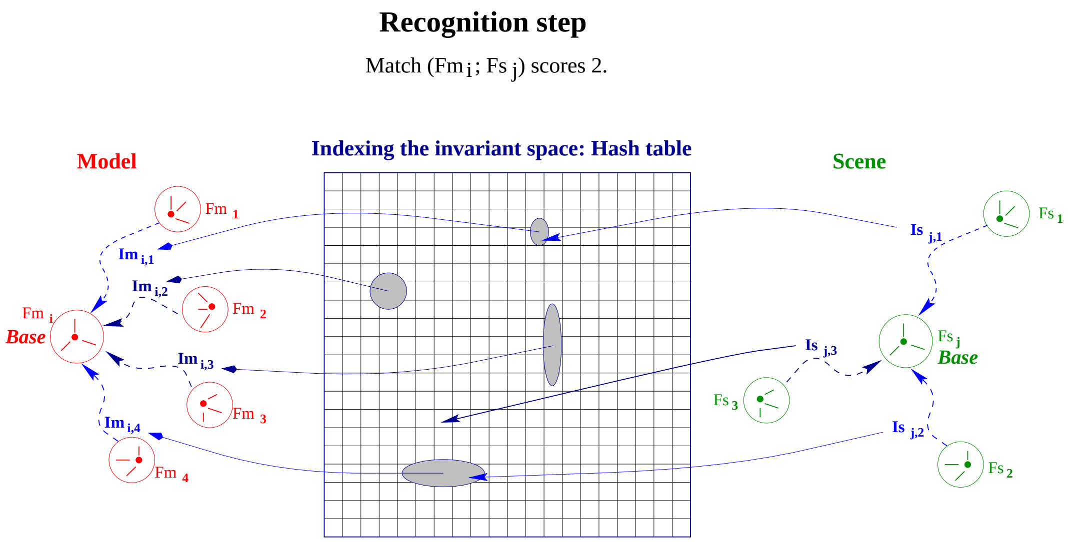


Figure 5: Recognition: for each scene frame couple, compute the 6D invariant vector and retrieve through the hash table every compatible model frame couple. For each such couple, tally a vote for the matching of the reference frames (here $Fm_i - Fs_j$).

list of the model reference frames found, and for each one accumulate the number of compatible couples. This will be the score for the matching of these model reference frames with the considered scene reference frame.

The process is repeated with each scene frame taken as the reference frame. The output is the list of model and scene matching reference frames with their associated score (see figure 5). 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).

The simple matching of a model and scene reference frames is sufficient for computing afterward a rigid transformation, but every compatible couple can brings up some additional information (the matching of secondary frames) that can be used to refine it at a small cost during this process, using for example an extended Kalman filter.

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 frames (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 the radius of the inscribed circle to the triangle used to form the reference frame, and d be the distance between 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 $\epsilon_{gh} = \epsilon\left(1 + \frac{d}{3r}\right)$. A strong supplementary constraint is given by the bound $\epsilon_d = 2\,\epsilon$ on the distance d. Concerning the 3 angular values (second part of the 6D vector), a theoretical error bound on each angle is $\epsilon_\theta = \frac{2\,\epsilon}{l} + O(\epsilon^2)$ 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 discretized, we verify at recognition time that the couples found have compatible invariant vectors.

As long as the hash table is sparsely distributed, this bounded error model for geometric hashing is sufficient. When it comes to very big proteins or big data bases, a probabilistic scheme [Rig92, RH93] would be more adapted.

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 on the translation vector and σ_θ^2 for error on the rotation vector. In our experiments we generally use $\sigma_d = 0.5$Å and $\sigma_\theta = 0.2$ rad.

- *Clustering*
  We have to aggregate the redundant information 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 each different match as follows:

  - Decide with a χ^2 test on Mahalanobis 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 each match be 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, a few tens at the very most.

- *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 are 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 (see figure 6).

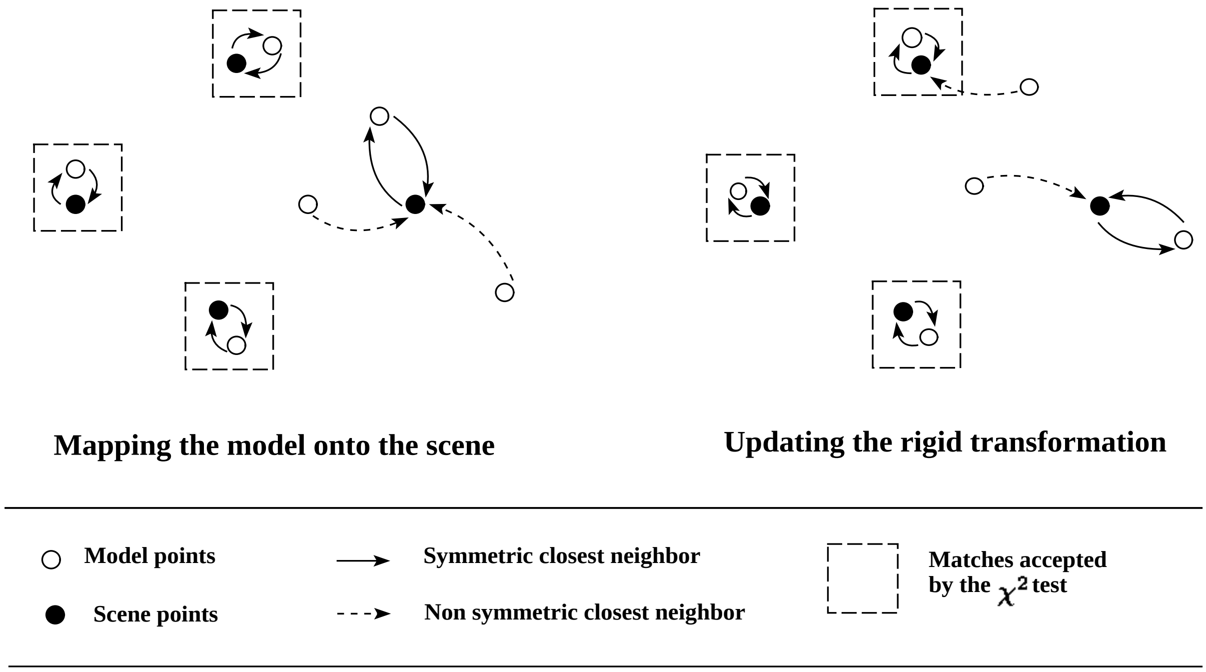


Figure 6: Left: the model is mapped onto the scene and three matches are satisfying the symmetric closest point and χ^2 test constraints. Right: these three matches are used to recompute the rigid transformation and update the position of the model. Since the same matches are satisfying the constraints, the algorithm can stop.

– 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 Mahalanobis 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.

This process is repeated until convergence (stability of matches) or a maximum number of iterations (usually about 10). 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 $O(n.\log n + n.k)$.

3.3 Algorithm analysis

- *Simplifications* : The following heuristics can help speeding up the geometric hashing step:

– 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 $O(M.m^2)$ for the preprocessing stage (comprising the k-D trees preprocessing of models). The whole recognition stage is in $O(n^2 + M.k.n)$. 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_c (clustering) and χ_v (verification) 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 [BKW+77, ABB+87]. Visualization is done using the RasMol program of R. Sayle [SB92]. For rigid transformations, we provide the translation in angströms and the rotation vector in radian. Experiments were done with and without the distance constraint of section (3.3) 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 [LZS+88]) and phage 434 CRO (PDB code 2CRO [MWH89]), whose Helix-Turn-Helix sequence are known to be [BM89, HA90] :

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

In this experiment, we choose the whole protein 2CRO as the model and look for common substructures with 2WRP: the two corresponding sequences are correctly matched, and 4 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
                                                                     ---------------

Cluster 2 : score 21
[ ............. ]
```

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 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.

4.1.1 Improvements of the algorithm

The last step (of the algorithm (the extension) is an alignment test realized as an improved iterative closest point [BM92] : the matching step looks for the symmetric closest point and the registration step discards outliers using a χ^2 test and recomputes the transformation. These two steps were previously realized only with points : we keep the matching part on points (we would like to be able to do it directly on frames), but we use the outlier rejection and registration scheme on frames (described in [PT96]) using a fixed noise model that determines the type of common substructures the algorithm extracts : a small noise only keep a small but very accurate set of matches, whereas a large noise authorizes more uncertain matches, and thus favours bigger and more global substructures matches.

Looking for a 3D binding motif, we used here a quite small model of isotropic noise ($\sigma_\theta = 0.1$ rad $= 5$ deg and $\sigma_d = 0.35$ Å). The algorithm ends up with only the correct matches from (15 MET - 66 MET) to (36 ALA - 87 ASN), discarding the four supplementary matches because they have different orientation. The last match (37 GLY - 88 SER) is also eliminated, and is indeed very arguable considering the distance after registration and especially the difference in orientation. The typical object precision due to the registration (on the 22 matched amino-acids) is given to $\sigma_{obj} = 0.29$ Å. We show in figure (7) the two proteins and in figure (8) the registration found. The only two other common substructures found score now 13 and 8 matches and correspond to alpha helices, which are very stable secondary structure elements.

In order to test the stability of our algorithm, we also did the experiment with a noise two times larger ($\sigma_\theta = 10$ deg and $\sigma_d = 0.7$ Å). We just find four additional matches preceding the beginning of the HTH motif, from (7 LYS - 61 LEU) to (11 ILE - 64 GLY). The two other clusters now score 14 and 9 matches.

This shows that the plain use of frames improves the robustness of motifs detection (the previous algorithm using only points was far more sensitive to the noise model). Indeed, the orientation of an amino acid is crucial to determine the position of collateral chains and most interactions of the protein happen within these side atoms. The position of these atoms is then not only determined by the position of the backbone but also its orientation and using just points to represent amino acids generally leads to a significant amount of additional matches with non compatible orientations. This implies a drastic reduction of selectivity for the matching process. In this case, frames bring rather more selectivity than more precision.

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 [Pa73], 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).

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 reported 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 appears to be very robust.

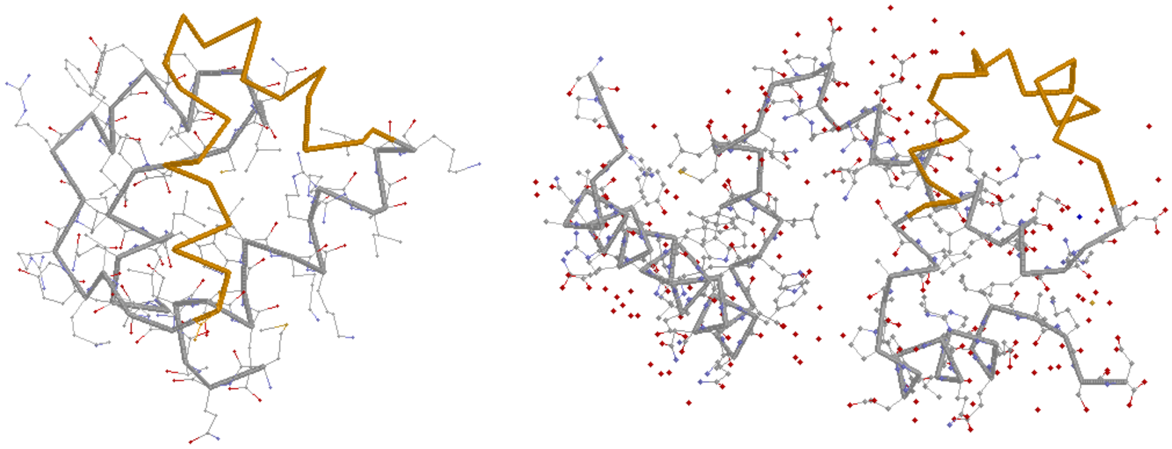


Figure 7: The CRO protein (2CRO) of phage 434 on the left and the tryptophan repressor of E. Coli (2WRP) on the right. The matched part (the HTH motif) is displayed in black.

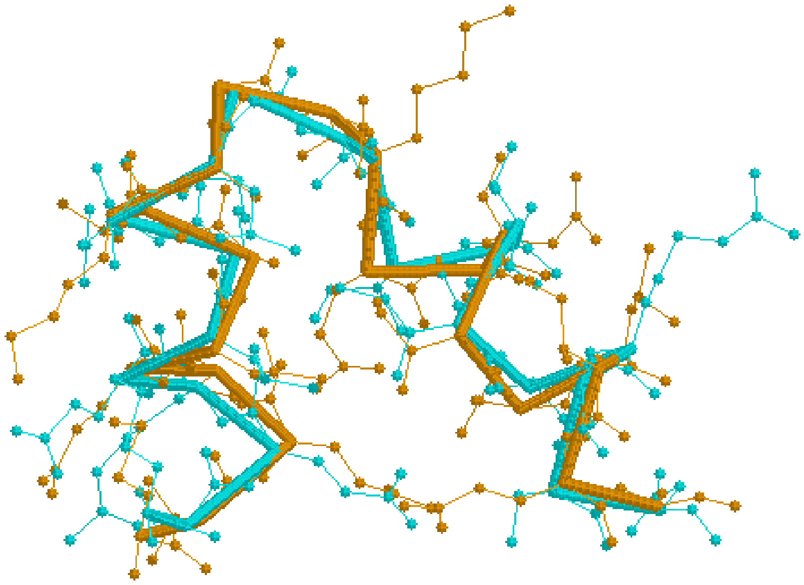


Figure 8: Registration found between the HTH motifs of 2CRO and 2WRP. We can see that not only the backbone is very well matched, but also collateral chains are pretty well conserved.

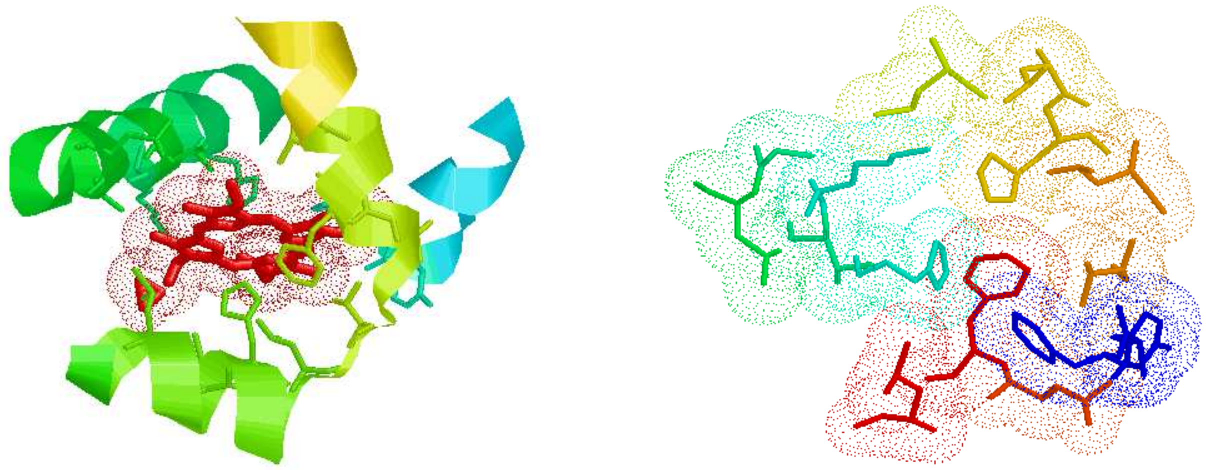


Figure 9: Left: the context of the heme in the α chain of human hemoglobin (4HHB). Right: the binding motif as defined in [LC80].

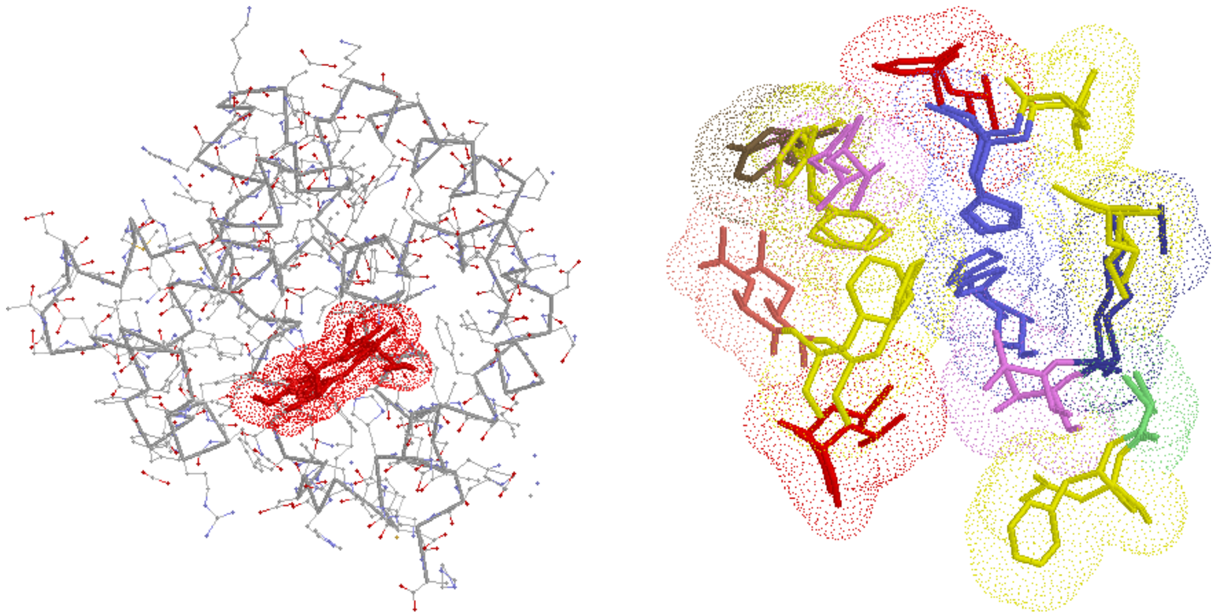


Figure 10: Example of protein structures binding the heme : the β chain of human hemoglobin (left) and its registration with the motif (right).

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.

4.3 Accuracy of frames

With the result of these two experiments, we can study the accuracy of our modeling with respect to the variability of amino acids. In figure (11) we draw the distribution of distance and angular error between matched frames in 2CRO and 2WRP (HTH experiment).

We can see that if two matched amino acids can be as far as 1.2 Å, their orientation differs from a maximum of 25^o, with a RMS of 16^o. As far as the heme experiments are concerned, the maximum and RMS distances are generally larger, but the angular error does not grow (table 3).

		4HHB	2DHB α	2DHB β	4MBN	2LHB
Distance	RMS	0.9	0.3	0.8	0.75	0.8
	Max	1.3	0.6	1.2	1.4	1.3
Angular error	RMS	$\pi/13$	$\pi/12$	$\pi/10$	$\pi/11$	$\pi/13$
	Max	$\pi/9$	$\pi/6$	$\pi/6$	$\pi/5$	$\pi/6$

Table 3: Error values for the heme experiment.

This study mainly shows that the orientation of amino acids is at least as stable as their position in a motif. This was in fact predictable since the orientation of an amino acid

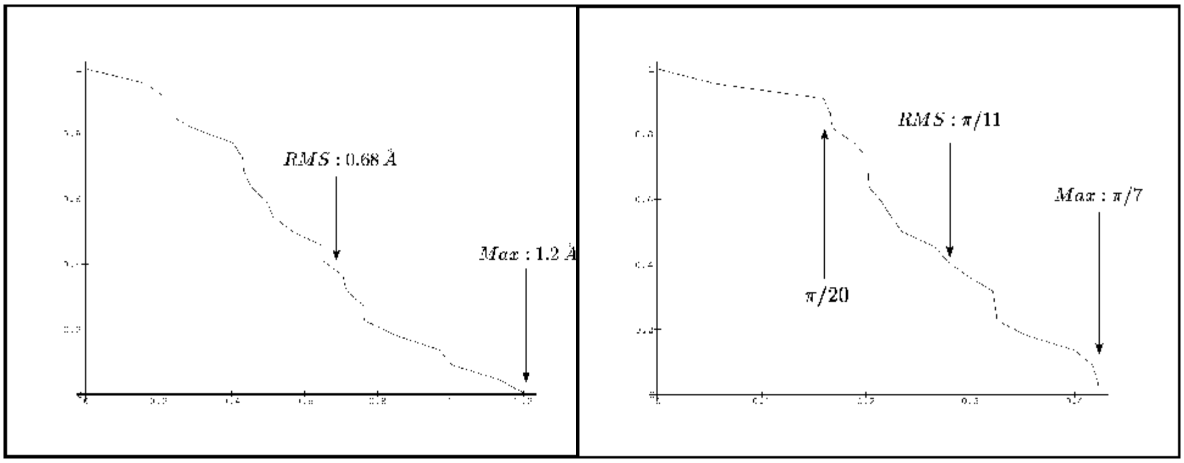


Figure 11: Distribution of distance (left) and angular error (right) between matched amino acids in 2CRO and 2WRP.

determines in a certain way the position of its residue. Our modeling with frames is then justified and allows to use this reliable information.

5 Conclusion

Modeling amino acids by the 3 atoms of their backbone allows to define a complete and unique associated reference frame, which turns out to be very stable. Each 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.

This algorithm is also currently working for substructure matching in volume images (medical images) with frames extracted from surfaces (extremal points). This stresses the analogy between 3D matching problems and points out the fact that frames can, in numerous cases, advantageously replace points.

Future work will be articulated upon three axes. We plan first to use a probabilistic scheme for geometric hashing, for instance [RH93]. A second direction would be the extension of our algorithm for multiple alignments research and 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

[ABB+87] 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, Bonn/Cambridge/Chester, 1987. Data Commission of the Int. Union of Crystallography.

[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*, October 1993. Also as INRIA Research Report No 2050.

[BKW+77] 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.

[BM92] P.J. Besl and N. McKay. A method for registration of 3-d shapes. *PAMI*, 14(2):239–256, 1992.

[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 - LNCS 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.

[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, Mai 1992.

[GA93] A. Guéziec and N. Ayache. New develoments on geometric hashing for matching 3d curves. In *Proceedings of Int. Conf on Comput. Vis. and Pat. Recog*, pages 703–704. IEEE Computer Society Press, June 1993.

[Gri90] W.E.L. Grimson. *Object Recognition by Computer - The role of Geometric Constraints.* MIT Press, 1990.

[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 efficient model-based recognition scheme. In *Proc. of Second ICCV*, pages 238–289, 1988.

[LZS+88] 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.

[Pa73] M.F. Perutz and al. Private communication, 1973.

[Pen93] X. Pennec. Correctness and robustness of 3D rigid matching with bounded sensor error. Research report 2111, INRIA, November 1993.

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

[PT96] X. Pennec and J.P. Thirion. A framework for uncertainty and validation of 3D registration methods based on points and frames. *Int. Journal of Computer Vision*, to appear, 1996.

[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, June 1993.

[Rig92] I Rigoutsos. *Massively Parallel Baysian Object Recognition.* PhD thesis, New York University, August 1992.

[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 and 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, Mai 1993. accepted for publication in IJCV.

[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, April 1990. Lecture Notes in Computer Science 427.