Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 10495–10499, December 1991
Biophysics

Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques

(three-dimensional structural comparison/crystallographic coordinates/efficient computer vision algorithm/macromolecular structure analysis)

RUTH NUSSINOV*† AND HAIM J. WOLFSON‡§

*Sackler Institute of Molecular Medicine, Faculty of Medicine and ‡Computer Science Department, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978 Israel; †Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Frederick Cancer Research Facility, Building 469, Room 151, Frederick, MD 21702; and §Robotics Research Laboratory, Courant Institute of Mathematical Sciences, New York University, 715 Broadway, 12th Floor, New York, NY 10003

Communicated by Jacob T. Schwartz, July 29, 1991 (received for review February 1990)

ABSTRACT **Macromolecules carrying biological information often consist of independent modules containing recurring structural motifs. Detection of a specific structural motif within a protein (or DNA) aids in elucidating the role played by the protein (DNA element) and the mechanism of its operation. The number of crystallographically known structures at high resolution is increasing very rapidly. Yet, comparison of three-dimensional structures is a laborious time-consuming procedure that typically requires a manual phase. To date, there is no fast automated procedure for structural comparisons. We present an efficient $O(n^3)$ worst case time complexity algorithm for achieving such a goal (where n is the number of atoms in the examined structure). The method is truly three-dimensional, sequence-order-independent, and thus insensitive to gaps, insertions, or deletions. This algorithm is based on the geometric hashing paradigm, which was originally developed for object**

isons are thus central to molecular biology. The problem we are faced with is to devise efficient techniques for routine scanning of structural data bases and searching for recurrences of inexact structural motifs. The degree of allowed errors is to be determined by the user.

The most commonly used computerized macromolecule comparison approaches deal mainly with comparison of the primary structure of molecules. They are based on character string comparison algorithms, most of which use variations of the dynamic programming technique (for a good survey, see ref. 13). Structural comparison is superior to this primary sequence analysis, since it takes into account the spatial geometric structure of the molecules involved and not only their order on the primary chain. The increasing need for direct structural analysis of macromolecules has led to the development of several computerized methods (14–16).

The mathematical problem stated above is closely related to the model-based recognition problem of 3D rigid objects. This problem has been intensively investigated in computer vision. One of the major problems in this field is to discover previously known objects in scenes, where some of the objects might appear to partially occlude each other. This is the, so called, model-based object recognition task (for extensive surveys, see refs. 17 and 18). By considering a molecule as an object consisting of many rigidly connected features (atoms), one can apply some of the computer vision techniques to our problem. Partial occlusion here is equivalent to the absence of partial substructures.

Several techniques have been suggested to tackle this problem. Some of them (19) exploit specific visual features that do not translate favorably to our problem. Others (20) employ tree search techniques resulting in exponential algorithm complexity. The most relevant techniques for our purpose are those known as alignment (21), pose-clustering (22), and geometric hashing (23) (for a comparison of these techniques, see ref. 24).

Recently, the geometric hashing paradigm for model-based object recognition was introduced by Lamdan *et al.* (23, 25, 26). This technique is especially geared toward recognition of partially occluded objects belonging to large-object data bases, and its complexity is a low-degree polynomial in the objects size. It is also very well suited for massive parallel implementation, and prototypes of this algorithm have been implemented on the highly parallel connection machine (27, 28). Techniques derived from computer vision have not been yet applied to molecular biology. We believe that their application will result in a significantly better performance than the manual graphics methods currently used not only

cient algorithms were developed for recognition of rigid objects both in two and three dimensions.

We present here a variant of the geometric hashing technique for recognition of identical partial structures in rigid 3D objects. For the moment we will use purely geometric language whose biological equivalents are as follows. A (geometric) rigid object is analogous to a molecule. Such an object consists of a set of points, which correspond to atoms. Each point may have a label (the name of an atom). Given a data base of known objects (molecules) and an observed object, the algorithm finds those objects in the data base, having large substructures nearly identical with substructures of the observed object. The points of matched substructures should have equivalent labels and identical 3D coordinates modulo translation and rotation (rigid motion) in space. No *a priori* knowledge of the desired substructure is assumed.

In a model-based object recognition system, one has to address two major interrelated problems, namely, object representation and matching. The representation used must be rich enough to allow reliable distinction between the different objects in the data base, yet terse enough to enable efficient matching. A major factor in a reliable representation scheme is its ability to deal with recognition of partial substructures. In the geometric hashing technique objects are represented as sets of geometric features (in our case, points), and their geometric relations are encoded using minimal sets of such features under the allowed object transformations (in our case, rigid motion). This is achieved by standard methods of analytic geometry invoking coordinate frames based on a minimal number of features and representing other features

Improvements of the Basic Paradigm. In the previous section we have described the basic geometric hashing scheme for 3D substructure detection. Various improvements are possible. In particular, one can design an $O(n^3)$ worst case algorithm for that purpose, although the practical run time of the previous version should also be much less than its worst case estimate. This other version is also more space efficient and requires a hash table of $O(n^3)$ only. On the other hand, we use somewhat weaker geometric and labeling constraints. In this section we sketch this second more efficient (in the worst case) algorithm.

In the scheme described above, we used full 3D bases that were associated with three-point RSs. One may, however, use somewhat weaker information: namely, two-point RSs. Given a two-point RS, any other (noncollinear) point in the 3D space defines a plane with this RS. Compute the two-dimensional (2D) coordinates of this point in the above mentioned plane using a 2D orthonormal coordinate frame, which is associated with the RS (the first point is the origin, and the vector from it to the second point defines the x axis). The address to the hash table this time will be the labels and the length of the RS segment, and the label and the 2D coordinates (in the appropriate plane) of the point. Since this procedure is done for all reference pairs, the hash table will take $O(n^3)$ space.

The recognition stage will be similar to the previous version, only this time one has to pick a reference pair on the observed object instead of reference triplet. Hence the worst case complexity reduces to $O(n^3)$. Since weaker geometric and labeling constraints are applied in this version, one may

(*iii*) The HTH motif was located in several bacterial repressor proteins just as noted in the annotated protein data bank (PDB). In our experiments we have compared three transcriptional regulatory proteins known to contain the HTH motif: tryptophan repressor (PDB code, 2WRP), λ Cro (PDB code, 1CRO), and phage 434 Cro (PDB code, 2CRO). To give a flavor of our experimental results we describe this example in more detail.

In 1CRO, there are four crystallographically unrelated monomers in the asymmetric unit. These monomers have been assigned chain identifiers O, A, B, and C. The dimer of 1CRO that exists in solution is presumed to be the O–B dimer, which is thought to be the one that actually binds DNA. We use the B monomer in the comparisons shown below, but comparisons using all four domains produce similar matches.

The sequence positions where the HTH motifs appear are as follows:

Protein	Positions	Sequence
2WRP	66–88	MS QRELKNELGA GIATITRGSNS
1CRO	14–36	FG QTKTAKDLGV YQSAINKAIHA
2CRO	15–37	MT QTELATKAGV KQQSIQLIEAG

In the three pairwise comparisons below (see Table 1), our method succeeds in matching the HTH motif from one protein to the HTH motif from the other. Very few other atom pairs are matched, showing that the only equivalent substructure between the proteins is the HTH motif itself. The atom pairs outside the HTH motif are 3D nonlinear matches.

For each pair of matching substructures Table 1 gives the sequence numbers of the matching atoms, the transformation between the substructures (translation parameters in ang-

Table 1. Pairwise matchings of the HTH motif in three proteins: 2CRO (phage 434), 1CRO (λ phage), and 2WRP (tryptophan repressor)

Model 2CRO	Scene 2WRP	Model 2WRP	Scene 1CROB	Model 1CROB	Scene 2CRO
				55-V	60-Q
					—
	—			44-I	53-N
63-T	103-V				—
	—			51-Y	50-M
				52-A	49-A
	—		—		—
37-G	88-S	88-S	36-A	36-A	37-G
36-G	87-N	87-N	35-H	35-H	36-A
35-E	86-S	86-S	34-I	34-I	35-E
34-I	85-G	85-G	33-A	33-A	34-I
33-L	84-R	84-R	32-K	32-K	33-L
32-Q	83-T	83-T	31-N	31-N	32-Q
31-I	82-I	82-I	30-I	30-I	31-I
30-S	81-T	81-T	29-A	29-A	30-S
29-Q	80-A		28-S	28-S	29-Q
28-Q	79-I	79-I	27-Q	27-Q	28-Q
27-K	78-G	78-G	26-Y	26-Y	27-K
26-V	77-A	77-A	25-V	25-V	26-V
25-G	76-G	76-G	24-G	24-G	25-G
24-A	75-L	75-L	23-L	23-L	24-A
23-K	74-E	74-E	22-D	22-D	23-K
22-T	73-N	73-N	21-K	21-K	22-T
21-A	72-K	72-K	20-A	20-A	21-A
20-L	71-L	71-L	19-T	19-T	20-L
19-E	70-E	70-E	18-K	18-K	19-E
18-T	69-R	69-R	17-T	17-T	18-T
17-Q	68-Q	68-Q	16-Q	16-Q	17-Q
16-T	67-S	67-S	15-G	15-G	16-T
	66-M	66-M	14-F	14-F	15-M
13-L	65-E	65-E	13-R	13-R	14-K
	64-G	64-G	12-M		13-I
	62-L	60-E	10-Y		11-I
	61-L	63-R	9-D	10-Y	10-R
	60-E	61-L	8-K		9-R
	59-E		7-L		8-K
	58-V	59-E	6-T	8-K	7-K
	57-I			7-L	6-L
	—		—		—
44-F	53-T			39-K	2-L

2. Pabo, C. O. & Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53**, 293–321.
3. Klug, A. & Rhodes, D. (1987) *Trends Biochem. Sci.* **12**, 464–469.
4. Gehring, W. J. (1987) *Science* **236**, 1245–1252.
5. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1988) *Science* **240**, 1759–1764.
6. Murre, C., McCaw, P. S. & Baltimore, D. (1989) *Cell* **56**, 777–783.
7. Suzuki, M. (1989) *EMBO J.* **8**, 797–804.
8. Mermod, N., O'Neill, E. A., Kelly, T. J. & Tjian, R. (1989) *Cell* **58**, 741–753.
9. Courey, A. J. & Tjian, R. (1988) *Cell* **55**, 887–898.
10. Tanaka, I., Appelt, K., Dijk, J., White, S. W. & Wilson, K. S. (1984) *Nature (London)* **310**, 376–381.
11. Rafferty, J. B., Somers, W. S., Saint-Girons, I. & Phillips, S. E. V. (1989) *Nature (London)* **341**, 705–710.
12. Abel, T. & Maniatis, T. (1989) *Nature (London)* **341**, 24–25.
13. Sankoff, D. & Kruskal, J. B. (1983) *Time Warps, String Edits and Macromolecules* (Addison-Wesley, Reading, MA).
14. Mitchel, E. M., Artymiuk, P. J., Rice, D. W. & Willet, P. (1989) *J. Mol. Biol.* **212**, 151–166.
15. Richards, F. M. & Kundrot, C. E. (1988) *Protein Struct.* **3**, 71–84.
16. Abagyan, R. A. & Maiorov, N. V. (1988) *J. Biomol. Struct. Dyn.* **5**, 1267–1279.
17. Besl, P. J. & Jain, R. C. (1985) *ACM Comput. Surv.* **17**, 75–154.
18. Chin, R. T. & Dyer, C. R. (1986) *ACM Comput. Surv.* **18**, 67–108.
19. Bolles, R. C. & Horaud, P. (1986) *Int. Robotics Res.* **5**, 3–26.
20. Grimson, W. E. & Lozano-Pérez, T. (1987) *IEEE Trans. Pattern Anal. Machine Intelligence* **9**, 469–482.
21. Huttenlocher, D. P. & Ullman, S. (1988) *Proceedings of the DARPA Image Understanding Workshop* (Morgan Kaufmann, San Mateo, CA), pp. 1114–1122.
22. Stockman, G. (1987) *J. Comput. Vision, Graphics, Image Process.* **40**, 361–387.
23. Lamdan, Y., Schwartz, J. T. & Wolfson, H. J. (1988) *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, New York), pp. 1407–1413.
24. Wolfson, H. J. (1990) in *Proceedings of the European Conference on Computer Vision*, ed. Faugeras, O. (Springer, Berlin), pp. 526–536.
25. Lamdan, Y. & Wolfson, H. J. (1988) *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, New York), pp. 238–249.
26. Lamdan, Y., Schwartz, J. T. & Wolfson, H. J. (1990) *IEEE Trans. Robotics Automation* **6**, 578–589.
28. Rigoutsos, I. & Hummel, R. A. (1991) *IEEE Workshop on Directions in Automated CAD-Based Vision* (IEEE, New York), pp. 76–84.
29. Thornton, J. M. & Gardner, S. P. (1989) *Trends Biochem. Sci.* **14**, 300–304.
30. Sutcliffe, M. J., Hanaeef, I., Carney, D. & Blundell, T. L. (1987) *Protein Eng.* **1**, 377–384.
31. Sarai, A., Mazur, R., Nussinov, R. & Jernigan, R. L. (1988)