New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design

R. Nilakantan,* N. Bauman, and R. Venkataraghavan

Lederle Laboratories, American Cyanamid Company, Pearl River, New York 10965

Received June 19, 1992

We present a method for the rapid quantitative shape match between two molecules or a molecule and a template, using atom triplets as descriptors. This technique can be used either as a rapid screen preceding the computationally expensive shape-based docking method developed by Kuntz and co-workers[1,2] or as a stand-alone method to rank compounds in a large database for their fit to a shape template. The merits and limitations of this method are discussed in detail with examples.

INTRODUCTION

It is recognized that biologically active molecules often depend for their action on a specific interaction with a receptor; this usually calls for some degree of shape complementarity between the ligand and the receptor. Several years ago, Kuntz et al.[1] presented a technique for characterizing possible binding sites of receptors (typically concavities on the solvent accessible surface[3] of the receptor) by sets of spheres. Such shape characterizations were subsequently used to find candidate ligands to fit the receptor on the basis of shape complementarity[2] and to screen large databases for them.[4-8]

Briefly, their method consists of docking the molecule onto a set of spheres (representing the negative image of a receptor site) by matching two edge-weighted molecular graphs derived from the interatomic distance matrixes of the candidate molecule and the spheres and scoring the goodness of fit. The scoring is done by giving points for favorable ligand-receptor contacts and penalties for bad contacts. Since graph matching by exhaustive search is computationally intense, they use heuristics to select some of the possible relative orientations of the two graphs and then test each orientation in turn. In their most recent work,[8] they use the docking graph technique[9,10] to carry out the graph matching. They have also described methods for speeding up the scoring step. These improvements reduce the overall search time as compared to their earlier methods.

Several variants of the basic method are used; for instance, if the structure of a receptor-inhibitor complex is known, it is possible to dock the candidate molecules onto the inhibitor and then score the fit. On the other hand, if the receptor alone is known, one can generate a set of spheres in the active site (as described above) to represent the presumptive ligand and match candidate molecules onto that image, scoring as above. A third possibility is that only a set of ligands is known; in such a case, it is possible to superimpose the ligands and generate spheres on the outside of the solvent-accessible surface of the union of the docked ligands. One could then score candidates by docking them into the artifically generated receptor. Another variant is to generate a set of spheres on the internal surface of the union of the docked ligands and use these spheres as a representative of the ideal ligand.

All of these involve graph matching to dock each ligand onto the receptor in hundreds or thousands of orientations, thus making the method rather slow. Hence, such searches are not usually conducted on very large databases. The Kuntz group, for example, has reported using a database of only about 10 000 compounds.

Thus, there is a clear need for a rapid technique to characterize molecular shapes without having to examine hundreds or thousands of docking orientations. Such a method would simply score the shape-fit between two molecules without actually discovering their correct relative orientation.

METHOD

The method requires three-dimensional coordinates for the template molecule and all the candidate molecules. The database could be experimentally derived, as is the Cambridge Crystal File,[11] or calculated by using any of several methods available for generating approximate 3D coordinates from connection tables.[12,13] Three major steps are involved in determining a numerical measure of shape match between two molecules:

(a) Every possible triplet of atoms in each hydrogen-suppressed molecular graph is taken as a triangle, and the lengths of the sides are calculated. These lengths are integerized, put into canonical order, and packed into a single 32-bit integer, the coded triplet. Duplicates are eliminated. See Appendix I for further details.

(b) The triplets in the two molecules are compared and the number in common is found.

(c) A simple formula is used to calculate a score which lies between 0 and 1. We have tried two different formulas as shown below:

$$s = 2c/(nt + nc) \quad (1)$$

$$s = c/nt \quad (2)$$

where s is the score; c is the number of triplets in common; nt is the number of triplets in the template; and nc is the number of triplets in the candidate.

In order to compare a molecule with an entire database of candidates, it is necessary to generate triplets for each molecule in the database. It is wasteful to have to do this every time we compare a compound against the database. On the other hand, since each molecule typically generates between 300 and 500 triplets, storing all of them would take too much space. We have found a compromise solution which involves the use of triplet shape signatures, henceforth referred to simply as "signatures".

The set of triplets in each molecule is represented by a single compact signature 2048 bits long. The signatures of all the molecules in the database are generated once and stored. When a probe molecule is to be compared against the database, its signature is generated and compared with the stored signatures of all the database entries. If there is a high similarity between the signatures of the probe molecule and a database entry, we generate triplets for that entry and do a detailed comparison as explained earlier. Otherwise, we

0095-2338/93/1633-0079$04.00/0 © 1993 American Chemical Society

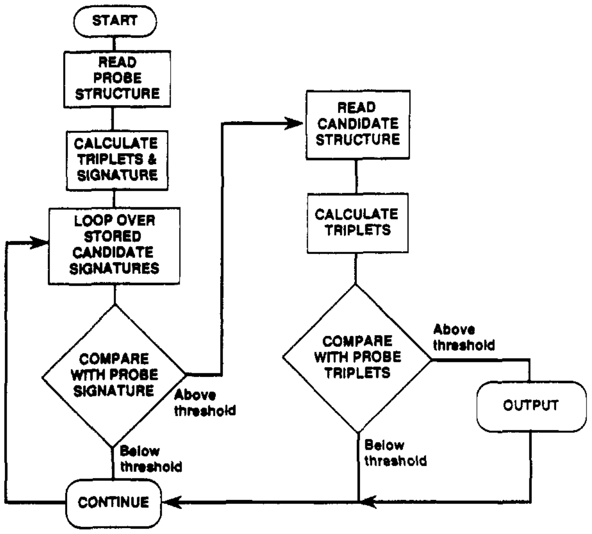


Figure 1. Flow diagram illustrating the steps in comparing a template molecule with a database of candidate molecules. The first step is to find the triplets in the template and then to use these to calculate a triplet signature for the template. Then the triplet signature of each candidate molecule is read and compared with that of the template. If the signature similarity is above a chosen threshold, the triplets are calculated for that candidate, and the triplet similarity between the template and the candidate is calculated. The similarity value and the candidate ID number are added to a rank list. If the signature similarity is below the chosen threshold, the candidate is skipped. After the entire database is processed in this manner, the rank list is sorted in descending triplet-similarity order.

skip that entry and move on to the next. Details of the signatures are given in Appendix II.

We have generated signature databases for the Cambridge Crystal File[11] as well as our proprietary 3D database of about 225 000 compounds (3D CL File).

A flow diagram showing the major steps in the search algorithm is shown in Figure 1.

RESULTS AND DISCUSSION

The triplet method can rapidly produce a numerical measure of shape match between two molecules. We present possible applications of this method with suitable examples.

Case 1. The method could be used to screen large databases to eliminate those candidates which have a low shape similarity with the template. One could then apply computer-based matching and docking techniques such as those of Kuntz and co-workers on those candidates that remain. We carried out an experiment to validate such a procedure.

Using the published X-ray coordinates of a complex of HIV-protease with the inhibitor MVT-101,[14] we carried out a shape search using a version of the docking program developed in collaboration with the Kuntz group.[4] The docked inhibitor was used as a template, and candidates were constrained to have between 15 and 40 atoms. We also required that there should be no bad intermolecular contacts between the docked candidate and the enzyme and also that at least one atom of the docked ligand should lie within 5 Å of the two active-site aspartates in the enzyme. Our test database consisted of 22 495 compounds derived from the Cambridge Crystal File (see Appendix II). We kept the 81 best-scoring compounds for further analysis.

We then also ran the triplet matching program using the same template against the database and ranked all the compounds in descending order of their scores. It should be pointed out that since this method does not dock candidates, it is not possible to impose intermolecular distance constraints such as given above. However, we did impose the constraint on the number of atoms. We then looked for the above-mentioned 81 shape search hits as we moved down the ranked list. Figure 2 shows a plot of the cumulative percent of shape search hits found versus the percentile rank of each compound in the list. It can be seen that 90% of the shape search hits appear in the first 29% of the rank-ordered list, representing a considerable enrichment. One could expect an even better enrichment if it were not for the receptor constraints used in the shape search. Indeed, in the next application, we present arguments for not using such constraints.

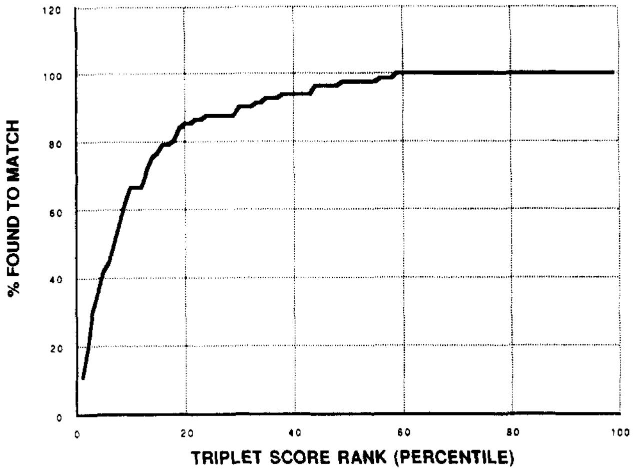


Figure 2. Plot of the cumulative percent of the 81 shape search actives found versus the percentile rank of each compound in the triplet-ranked list. Note that 90% of the 81 shape search selections are to be found in the top 29% of the triplet-ranked list.

Case 2. A second mode of using the method would be to eliminate the computer-based matching step (shape search) altogether, simply taking the highest scoring compounds (say the top 50) and docking them manually onto the template or the active site. The great advantage of such a procedure is that the human being has a far better pattern-matching capability than the computer. Furthermore, after manual docking, one can use simple molecular mechanics to relax the structure and eliminate trivial short contacts. (With automated docking, it often happens that good shape matches are discarded because of short contacts, which could have easily been eliminated by molecular mechanics.)

To explore this possibility we carried out some experiments. Grootenhuis et al.[15] have discussed the discovery of drugs that bind to the minor groove of DNA using the netropsin-B-DNA crystal structure[16] as the prototype. These authors generated a set of spheres in the minor groove representing the negative image of the shape of the groove and docked a series of candidate molecules onto them. They have published the top 10 structures (including netropsin) and point out that the best ranking structure (CCD refcode CCATAG10) is in fact a potent antitumor agent (CC-1065) known to interact with DNA.[17-19] We experimented with the triplet method using the bound netropsin structure as the template and picked out structures (from a database of 22 498 compounds, which included the 10 best scoring compounds obtained by Grootenhuis) showing a high degree of shape match.

Figure 3 shows eight of these structures and the ranks assigned to them by the triplet method. It is satisfying that the triplet method also assigns the top rank to CCATAG10. Further, of these eight, the compound ranked lowest by the triplet method (refcode COSHUN) has a rank of 1860 out of 22 498. Thus, all eight compounds appear in the top 8.3%, a remarkable enhancement. Examination of the structure of

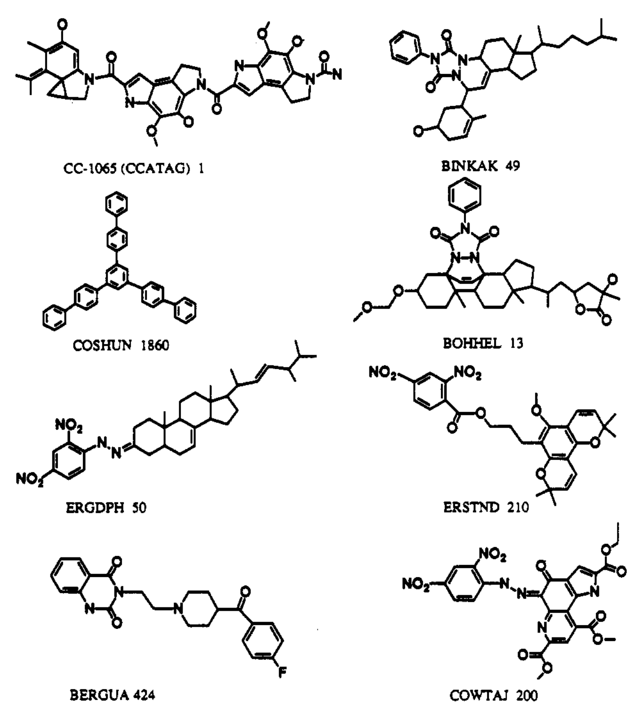


Figure 3. Eight highest ranked molecules given by Grootenhuis et al.[15] in their study of DNA groove-binding drugs. Alongside the refcode is given the ranking of each compound by the triplet method. Actually, in their publication, these authors have given the 10 highest ranking compounds. Netropsin is one of them, and we have not shown it since the DNA-bound netropsin molecule was itself the template in our triplet study. Another one of the top 10 (refcode DBZCTD) seems to be erroneously represented in the Cambridge Crystal File, and so we have chosen not to show it.

the lowest ranked compound shows that it really does not have much overall shape similarity to netropsin and, hence, the low score. The next lowest ranked compound (ERSTND) has a rank of 424 which corresponds to a percentile score of 1.2%. Thus seven out of eight compounds appear in the top 1.2% of the triplet-ranked list. This clearly illustrates the ability of the triplet method to identify compounds likely to be ranked high by the shape search method.

The nine compounds ranked highest by the triplet method are shown in Figure 4. We docked each of these onto the minor groove of B-DNA using the bound netropsin as a rough guide. No attempt was made to achieve any functional group complementarity since that is outside the scope of the methods being discussed. Of the ten, four appeared to be able to fit into the minor groove provided they were relaxed a little. We used the BIOGRAF molecular modeling package[20] and carried out a simple energy minimization on each of these structures keeping the DNA structure rigid. It was seen that all four fit the minor groove remarkably well. Figure 5 shows stereoviews of these docked structures. Of particular interest is the fact that each of these molecules, like netropsin, adopts a helical shape to complement the minor groove.

Case 3. A third mode of using the triplet method, particularly relevant to the drug industry, would be to eliminate the docking step altogether and simply screen the candidates ranked highest by the triplet method in the appropriate biological assay. This is a meaningful alternative because drug screening is typically quick and inexpensive, while docking compounds is not. Once we find compounds with biological activity, we could try to dock them to find a retrospective explanation for activity and complete the design-test-design cycle. This mode of operation would have discovered CC-1065 in the previous example.

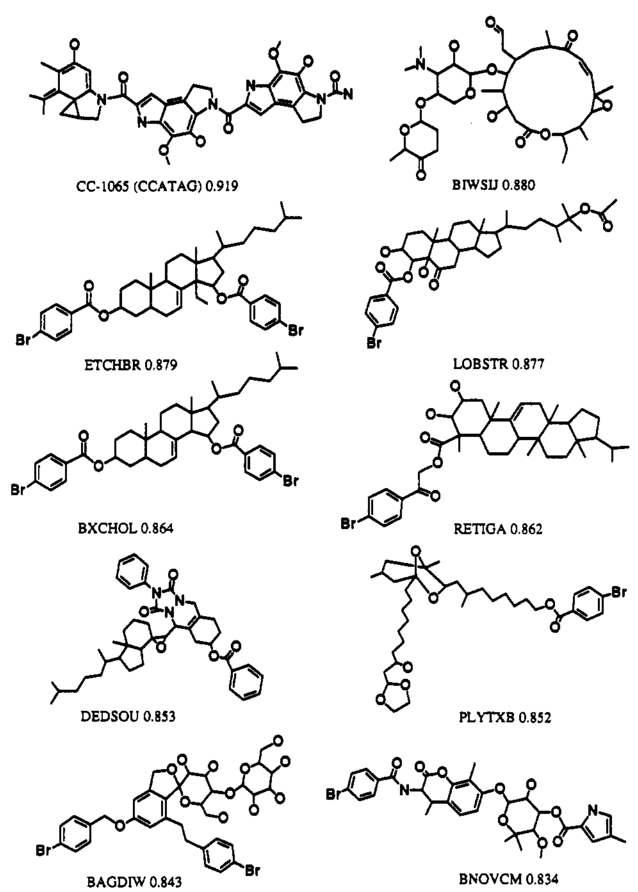
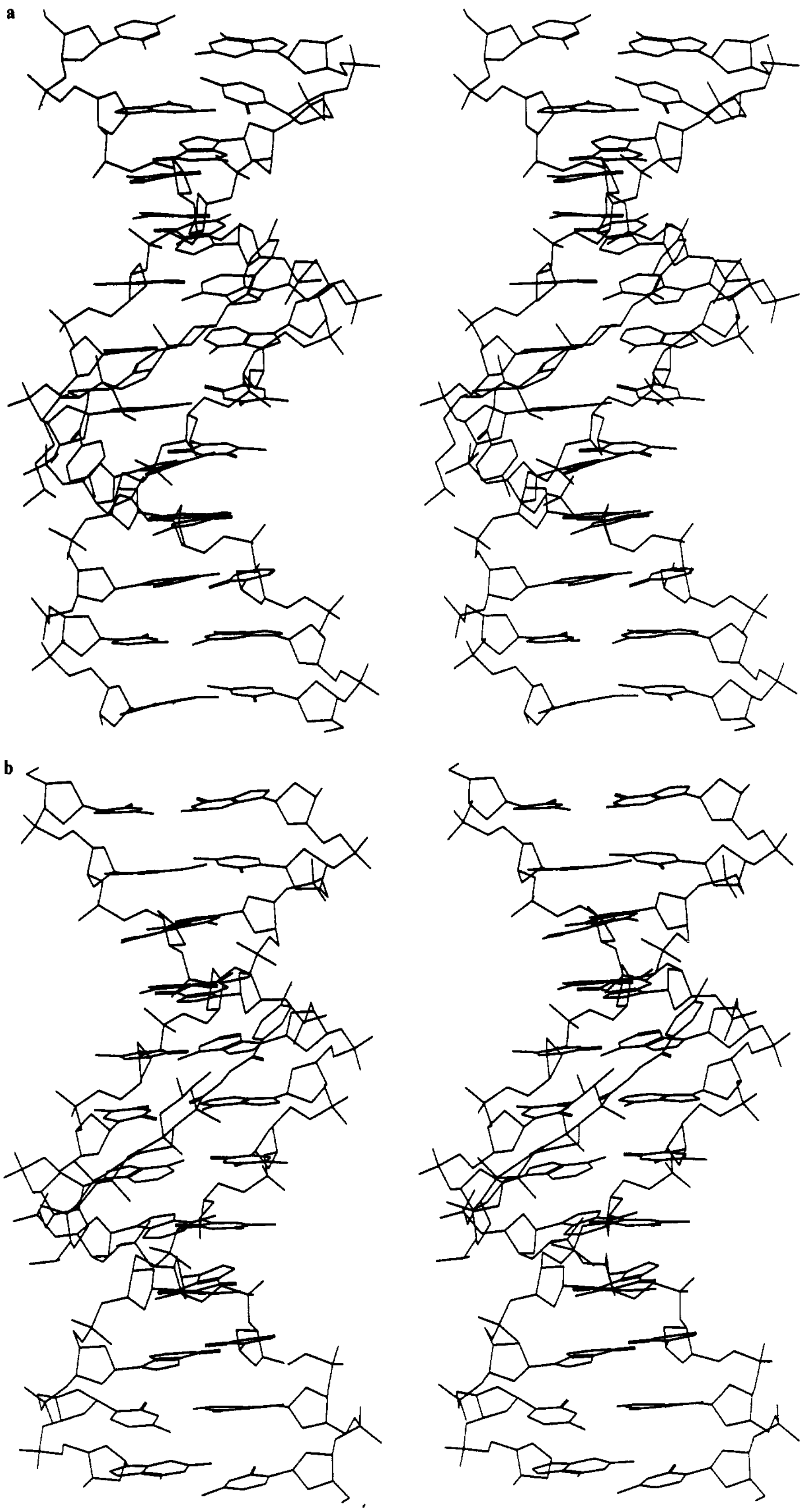


Figure 4. Ten molecules ranked highest by the triplet method for their shape similarity to the DNA-bound netropsin template. Alongside the refcode is given the triplet similarity score for each molecule.

Why triplets? The complete set of interatomic distances in a molecule *without any connectivity information* does not uniquely determine the molecule within rigid transformation. Although it is possible to generate 3D coordinates, using the distance geometry method,[21] for particular assignments of the distances to pairs of atoms, in general more than one set of coordinates is consistent with an interatomic distance set. The set of all triangles formed by a molecule, rather than the set of interatomic distances, obviously provides a more restrictive description, resulting in fewer spurious matches. By extension, all the quartets of atoms in the molecule would provide an even richer description. However, the number of quartets of atoms in a molecule would be much larger than the number of triplets. As our purpose is to use these descriptors to estimate shape similarity, an overly rich description would prevent us from detecting similarities in molecules with roughly similar shape. Our choice of triplets as shape descriptors represents a trade-off between obtaining a sufficient level of description and the available computational resources.

Even with the use of triplets, a molecule with 25 atoms generates 2300 triplets, a rather large number. So we adopted a simplification and eliminated duplicates. The experiments discussed above illustrate that in spite of this simplification, we are able to obtain an adequate description of molecular shape.

Scoring Scheme. As discussed earlier, we use two different scoring formulas. The first is the same as we use to calculate the topological similarity between two molecules and has been discussed elsewhere.[22,23] This formula corresponds to a symmetric match, penalizing a candidate molecule whenever



it has either too few target triplets or too many extraneous triplets. The second formula penalizes the candidate when it has too few target triplets but not when it has extraneous triplets. The latter corresponds roughly to matching the template with a substructure of the candidate. In the two studies reported here, we used the second scoring function.

Speed of the Method. The great speed of the present method is derived from the fact that each molecule or collection of points is represented by a set of descriptors encoded as integers; comparison of two molecular shapes requires only the comparison of two sorted lists. The price we pay is that the relative orientation of the two molecules is not discovered. The slowest steps are the one which detected the triplets (order n^3, where n is the number of atoms or points in the structure) and the one required to build a sorted list of descriptors (order n^3 log (n)). The use of triplet signatures as screens further speeds up the process, particularly when scanning large databases. On our VAX 8650 machine, a typical triplet search

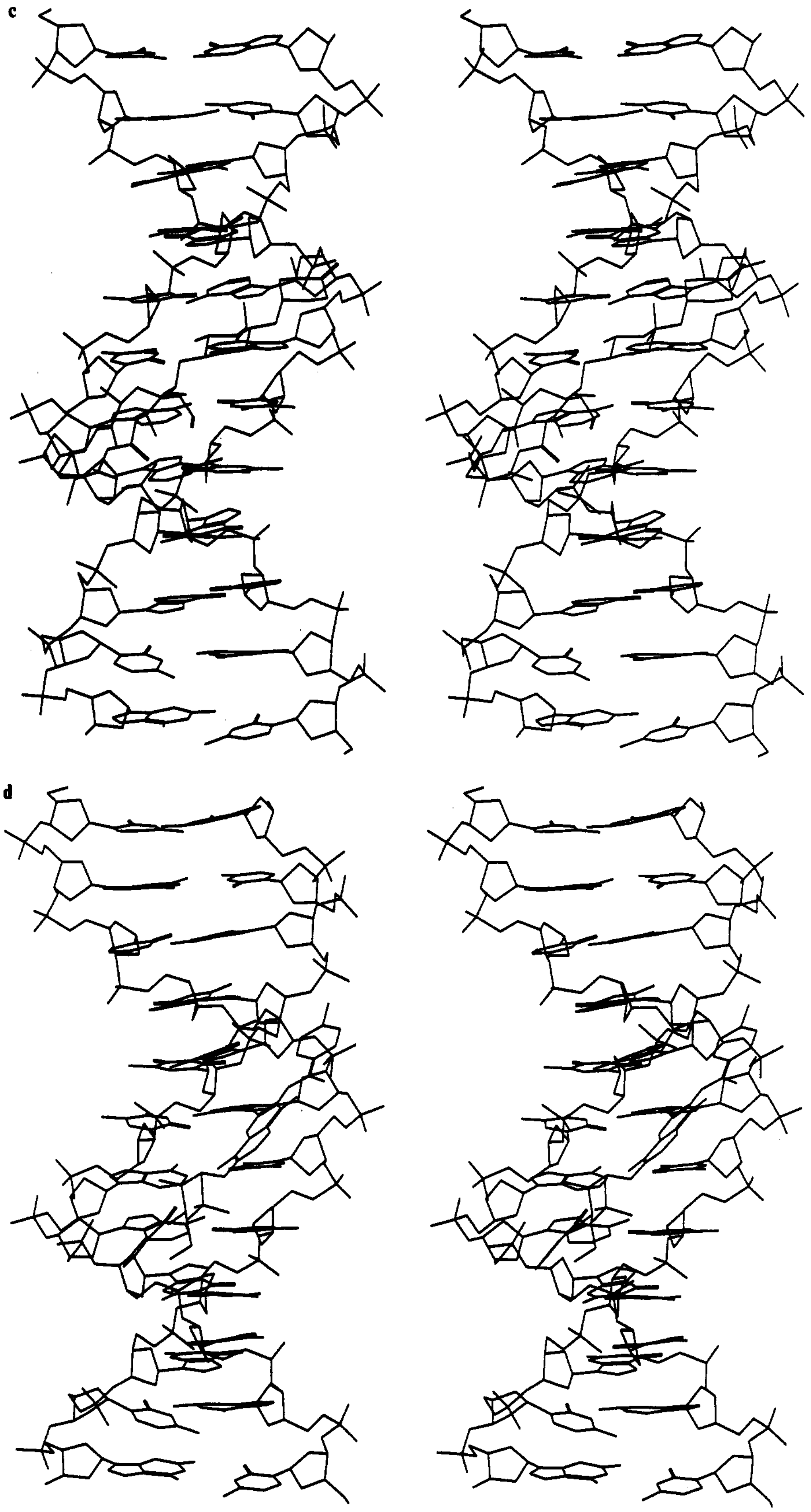


Figure 5. Stereo pairs of some compounds selected by the triplet method for their shape similarity of the DNA-bound netropsin template. The compounds have been manually docked into the DNA minor groove and have been minimized using molecular mechanics. Panels a and b are shown on the opposite page. (a) Refcode ETCHBR; (b) Refcode RETIGA; (c) Refcode BIWSIJ; (d) Refcode BNOVCM. Note the high shape complementarity between the minor groove and the bound molecules.

over our corporate database takes 2–3 h of CPU time. This enables us to carry out shape-based studies on large databases.

Advantages of the Method. (a) The use of triplets as a fast prescreen to the shape search methods of Kuntz and co-workers avoids costly graph matching and docking computations on a large fraction of the compounds in a database.

(b) In its stand-alone mode, the method rapidly picks out compounds with a high overall shape similarity to the probe compound. The examples of potential DNA groove-binding drugs show the ability of the method to perceive such subtle details as the overall helical shape of the bound netropsin molecule.

(c) Furthermore, in shape matching studies where a small set of spheres is used to represent a shape, there is inevitably the problem of instability; that is, a small change in the positions of these spheres could dramatically change the nature of the selected candidates. Use of the triplet method as an alternative to shape search overcomes some of these difficulties.

Disadvantages of the Method. The obvious disadvantage of the method is that it does not tell us the docking orientation between the candidate ligand and target receptor molecules. As a consequence, it is not possible to apply any active-site constraints on the molecule. However, as discussed above, this is probably not a serious drawback particularly in the context of drug screening.

Triplets with Chemical Characteristics. It would obviously be of interest to add chemical information to the atom triplets we have discussed. This, however, leads to an explosion in the number of triplet types, thus making the method unworkable. There are, however, certain simplifications we can use in the definition of triplets. We have tried the following:

1. Use only triplets with at least one heteroatom.
2. Disregard triplets with sides greater than 30 Å.
3. Use only 10 different atom types: C, N, O, P, S, F, Cl, Br, I, and Si.

These rules were chosen with the pharmaceutical screening application in mind. With these simplifications, we find that the method can be made to work at about the same speed as the purely geometrical method and with similar demands on memory and storage. We are currently experimenting with this variant.

CONCLUSION

The examples we have discussed above show that, for purposes of shape-based structure selection, the set of triplets obtained from a collection of atoms or points representing a shape template is an adequate representation of the overall shape of that collection.

APPENDIX I

Method of Calculating Triplets. Triplet descriptors are found by calculating the sides of the triangle formed by every set of 3 points in the set. We exclude distances of 100 Å or more. These lengths are then placed into a set of discrete bins. The bin number is calculated by multiplying the distance by 2 and retaining the integer part of the quotient. This is equivalent to using a bin size of 0.5 Å.

The three integers representing each triplet are then sorted into ascending order and packed into a single 32-bit integer using the following formula:

$$nt = n1 + 1000n2 + 1\,000\,000n3$$

where nt is the packed integer, and $n1$, $n2$, and $n3$ are the three integerized and sorted sides of the triplet.

From the packing formula, it is easy to see that we could have used finer bins, e.g., 0.125 Å, one-fourth of the present bin size. There are two reasons why we chose not to do so: (a) Reducing the bin size to 0.125 Å might make the method too sensitive to small changes in structure; we wanted a good characterization of overall shape and not fine details. (b) Reducing the bin size would increase the number of triplets in a molecule, leading to excessively long lists of numbers to be sorted. Since sorting is one of the rate-limiting steps, we kept it to a minimum.

APPENDIX II

Triplet Shape Signatures. The basic idea of the shape signature is to generate a single short bit pattern representing the entire set of triplets in a molecule. Each triplet in a database molecule is hashed to two numbers between 1 and 2048, and the corresponding pair of bits is set in a bit string. These bit strings are stored in a separate database and used as prescreens.

Since we are using only 2048 bits for our signatures, it is obvious that with 300–500 triplets in a typical molecule, our signatures cannot be unique representations. However, they turn out to be adequate as a screening device.

To compare a probe molecule with the database, we generate a signature for it and compare it to the stored signatures of each database molecule. We compute a symmetric similarity using the formula

$$s = 2c/(na + nb)$$

where s is similarity; c is the number of bits in common; na is the number of set bits in the probe signature; and nb is the number of set bits in the database molecule's signature.

Notice that this formula is identical to that used in computing the triplet similarity (see eq 1 in the Methods section). If the similarity value is greater than, say, 0.6, we generate triplets for the database entry and do a detailed triplet similarity calculation; otherwise we reject it. By choice of the cutoff, we can virtually eliminate the possibility that a database molecule with a high triplet similarity to the probe molecule is rejected. In our implementation, the program dynamically alters the cutoff so that a certain level of screening is maintained.

It should be remembered that with this highly reduced signature representation there is the risk that we miss compounds. In practice, however, this does not happen, particularly with compounds having high similarities to the probe.

APPENDIX III

Database of Compounds. The tests we have reported in this paper were all conducted on a set of compounds obtained from the Cambridge Crystal File. Some pre-processing was done to obtain this subset. First, we dropped all atoms other than C, N, O, F, Si, P, S, Cl, Br, and I in each molecule. Then, for the atoms that remained, we generated connectivity information from the interatomic distances by treating all atoms less than 2 Å apart as being bonded. We then represented the molecule as a graph, treating bonds as edges and atoms as nodes, and identified the largest connected component, using an algorithm due to Floyd.[24] If this component had between 10 and 50 atoms, we kept it; otherwise we discarded it.

We used the same set of compounds for the Kuntz-style shape search runs as well as the triplet-based calculations.

REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(2) Desjarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149–2153.

(3) Richards, F. M. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

(4) Desjarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722–729.

(5) Desjarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor-binding site of known three-dimensional structure. *J. Med. Chem.* **1989**, *31* (4), 722–729.

(6) Desjarlais, R. L.; Seibel, G. L.; Kuntz, I. D. Second-generation computer-assisted inhibitor design method. *ACS Symp. Ser.* **1989**, *413*, 60–69.

(7) Desjarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; Demontellano, P. R. O.; Decamp, D. L.; Babe, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus-1 protease. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87* (17), 6644–6648.

(8) Schoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13* (3), 380–397.

(9) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **1984**, *5* (1), 24–34.

(10) Smellie, A. S.; Crippen, G. M.; Richards, W. G. Fast drug-receptor mapping by site-directed distances: A novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (3), 386–392.

(11) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge crystallographic data center: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr.* **1979**, *B35*, 2331–2339.

(12) CONCORD: A program for the rapid generation of high quality approximate 3-dimensional molecular structure. The University of Texas, Austin, and Tripos Associates: St. Louis, MO, 1988.

(13) Dolata, D. P.; Leach, A. R.; Prout, C. K. WIZARD: AI in conformational analysis. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73–85.

(14) Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R. L.; Kent, S. B. H.; Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3-Å resolution. *Science* **1989**, *246*, 1149–1152.

(15) Grootenhuis, P. D. J.; Kollman, P. A.; Seibel, G. L.; Desjarlais, R. L.; Kuntz, I. D. Computerized selection of potential DNA binding compounds. *Anti-Cancer Drug Des.* **1990**, *5*, 237–242.

(16) Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. The molecular origin of DNA–drug specificity in netropsin and distamycin. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 1376–1380.

(17) Chidester, C. G.; Kreuger, W. C.; Misak, S. A.; Duchamp, D. J.; Martin, D. G. The structure of CC-1065, a potent antitumor agent, and its binding to DNA. *J. Am. Chem. Soc.* **1981**, *103*, 7629–7635.

(18) Hurley, L. H.; Reynolds, V. L.; Swenson, D. H.; Petzold, G. L.; Scahill, T. Reaction of the antitumor antibiotic CC-1065 with DNA: structure of a DNA adduct with DNA sequence specificity. *Science* **1984**, *226*, 843–846.

(19) Zimmer, C.; Waehnert, U. Noninteracting DNA-binding ligands: specificity of the interaction and their use as tools in biophysical, biochemical and biological investigations of the genetic material. *Prog. Biophys. Mol. Biol.* **1986**, *47*, 31–112.

(20) BIOGRAF is a product of Molecular Simulations Inc., Sunnyvale CA.

(21) Crippen, G. M. In *Distance Geometry and Conformational Calculations*; Research Studies Press: New York, 1981.

(22) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure–activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(23) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.

(24) Floyd, R. W. Algorithm 97 Shortest Path. *Commun. ACM* **1962**, *5*, 345.