J-CAMD 365

# An inequality for 3D database searching and its use in evaluating the treatment of conformational flexibility

John H. Van Drie

*Pharmacia & Upjohn Inc., Kalamazoo, MI 49001, U.S.A.*

Received 28 February 1996
Accepted 2 June 1996

*Keywords:* 3D database searching; Conformational flexibility

## Summary

A mathematical formula is introduced for predicting the number of hits that should be observed in a flexible 3D database search, based on the results of a set of related queries. The projected number of hits is always greater than or equal to the actual number of hits, the discrepancy being due to imperfect treatment of conformational flexibility of the molecules. Hence, the difference between the projected and actual number of hits, δ, serves to measure how well conformational flexibility is being treated, in a manner that is objective, easy for a user to quickly verify, and independent of the particular algorithm for flexible 3D database search. It is shown that δ is a function both of how well conformational flexibility is treated and of the precision of the query. When the distance constraint is defined only to a precision of ± 2.0 Å, in a single-conformer database of drug-like molecules δ values of only 0.03 are found, while in a single-conformer database of di- and tripeptides, δ is 0.15. At increased precision, a flexible 3D database search becomes critical. For a single-conformer database, using a query of precision ± 0.2 Å, applied to a database of drug-like molecules, δ is 0.97; applied to a database of di- and tripeptides, δ is 2.21. By contrast, treating conformational flexibility by storing up to 100 conformers per molecule, at this precision, applied to a database of drug-like molecules, δ is 0.002; applied to a database of di- and tripeptides, δ is 0.07. This inequality, and hence δ, is defined only for database queries containing a single distance constraint; how the inequality may generalize to higher-dimensional queries is still unclear.

## Introduction

Three-dimensional database searching is a class of computational methods for attempting to discover new biological activities of known compounds by searching a database of 3D structures to find molecules which satisfy a set of geometric and topological constraints. The first procedure for 3D database searching was introduced by Gund et al. [1]. In their approach, an atomic pharmacophore was used to represent those constraints. The later work of Willett and co-workers [2] and Sheridan et al. [3] is patterned after this approach. Kuntz et al. [4] introduced a procedure based on steric complementarity to a protein pocket as the basis for the constraints. VanDrie et al. [5] introduced a general approach in which pharmacophoric patterns, steric complementarity, or general geometric relationships could be used to define the 3D constraints. A number of other investigators followed this approach [6–9]. Alternatively, Bartlett et al. [10] used strictly vectorial relationships for the geometric constraints.

These early efforts primarily relied on searching a database of a single conformer per molecule, either experimentally determined (Cambridge Crystallographic Database) or computationally generated by CONCORD [11]. We initially described populating a database with multiple conformers per molecule as a method for treating conformational flexibility [5], but since these conformers were generated with molecular mechanics the large computational resources needed prevented its practical application for databases of hundreds of thousands of molecules. In the past few years, a number of approaches have been introduced for treating conformational flexibility in 3D database search, involving either computations on-the-fly [8,12] or special construction of the 3D database prior to the search [9,13].

The need exists for a way to objectively compare these different approaches for treating conformational flexibility, i.e. in a way that is not inherently biased toward one approach or another. A method is described here for evaluating the treatment of conformational flexibility

0920-654X/$ 6.00 + 1.00 © 1996 ESCOM Science Publishers B.V.

TABLE 1
AVERAGE NUMBERS OF CONFORMERS PER MOLECULE FOR DIFFERENT DATABASES THAT WERE CONSTRUCTED

| Database | Average number of conformers/molecule |
|---|---|
| NCIX1 | 1.0 |
| NCIX5 | 4.75 |
| NCIX10 | 8.77 |
| NCIX30 | 21.1 |
| NCIX100 | 42.3 |
| DitripepX1 | 1.0 |
| DitripepX5 | 4.992 |
| DitripepX10 | 9.948 |
| DitripepX30 | 29.46 |
| DitripepX100 | 70.15 |

which is objective and easily applied by any user to any pharmacophoric 3D database search system. This method leads to the insight that proper treatment of conformational flexibility is not an absolute requirement. Rather there is a complementarity between the precision of the query and the demand for the treatment of conformational flexibility: imprecise (unselective) queries can be treated effectively by a single-conformer database, while precise, selective queries demand a proper treatment of conformational flexibility.

There are a number of procedures for treating conformational flexibility in 3D database searching. Here only one procedure is investigated, namely conformationally flexible 3D searching of precalculated sets of discrete conformers of each molecule. It should be stressed, however, that the method of analysis employed here is general and can be applied to any method of conformationally flexible 3D search.

## Methods

In order to evaluate different approaches to treating conformational flexibility in 3D database search, a mathematical inequality which holds for the number of hits returned from a set of related queries may be described as follows. A query composed of a single distance constraint between two features A and B is constructed such that A and B are constrained to lie within the range $[d_{min},d_{max}]$ of distances. It is shown in the Appendix that the number of hits the query returns, $N[d_{min},d_{max}]$, must obey the following inequality:

$$N[d_{min},d_{max}] \leq N[0,d_{max}] + N[d_{min},\infty] - N[0,\infty] \qquad (1)$$

where $N[0,d_{max}]$, $N[d_{min},\infty]$, and $N[0,\infty]$ are the numbers of hits obtained with queries constraining the features A and B in the ranges $[0,d_{max}]$, $[d_{min},\infty]$, and $[0,\infty]$, respectively. Note that in actual calculations '∞' is effectively equivalent to 9999 Å. Equality holds only when conformational flexibility is treated properly. The set of quantities on the right-hand side of Eq. 1 is referred to as the *projected* number of hits within $[d_{min},d_{max}]$:

$$PN[d_{min},d_{max}] = N[0,d_{max}] + N[d_{min},\infty] - N[0,\infty] \qquad (2)$$

and the relative difference between $PN[d_{min},d_{max}]$ and $N[d_{min},d_{max}]$ is denoted as:

TABLE 2
OBSERVED AND PROJECTED NUMBERS OF HITS USING THE QUERY OF BASIC NITROGEN 6 ± η Å FROM A CARBOXYL AGAINST THE NCI DATABASE CONSTRUCTED WITH VARYING NUMBERS OF CONFORMERS/MOLECULE

| η (Å) | Database | $N[6-\eta, 6+\eta]$ | $N[0, 6+\eta]$ | $N[6-\eta, \infty]$ | $PN[6-\eta, 6+\eta]$ | $\delta[6-\eta, 6+\eta]$ |
|---|---|---|---|---|---|---|
| 2.0 | NCIX1 | 1436 | 2449 | 1681 | 1474 | 0.027 |
| 1.0 | NCIX1 | 814 | 2365 | 1190 | 899 | 0.10 |
| 0.5 | NCIX1 | 422 | 2286 | 981 | 611 | 0.45 |
| 0.25 | NCIX1 | 271 | 2256 | 933 | 533 | 0.97 |
| 2.0 | NCIX5 | 1633 | 2537 | 1782 | 1663 | 0.018 |
| 1.0 | NCIX5 | 1074 | 2461 | 1324 | 1129 | 0.051 |
| 0.5 | NCIX5 | 771 | 2387 | 1158 | 889 | 0.15 |
| 0.25 | NCIX5 | 514 | 2361 | 1111 | 816 | 0.59 |
| 2.0 | NCIX10 | 1682 | 2553 | 1791 | 1688 | 0.0036 |
| 1.0 | NCIX10 | 1151 | 2490 | 1352 | 1186 | 0.030 |
| 0.5 | NCIX10 | 865 | 2432 | 1172 | 948 | 0.10 |
| 0.25 | NCIX10 | 643 | 2406 | 1133 | 883 | 0.37 |
| 2.0 | NCIX30 | 1708 | 2571 | 1794 | 1709 | 0.0006 |
| 1.0 | NCIX30 | 1218 | 2518 | 1367 | 1229 | 0.010 |
| 0.5 | NCIX30 | 949 | 2469 | 1177 | 990 | 0.043 |
| 0.25 | NCIX30 | 786 | 2449 | 1146 | 939 | 0.19 |
| 2.0 | NCIX100 | 1772 | 2637 | 1793 | 1774 | 0.0011 |
| 1.0 | NCIX100 | 1308 | 2595 | 1372 | 1311 | 0.0023 |
| 0.5 | NCIX100 | 1052 | 2555 | 1178 | 1077 | 0.020 |
| 0.25 | NCIX100 | 908 | 2541 | 1153 | 1038 | 0.14 |

$N[0,\infty]$ is 2656 in each case. The data are plotted in Fig. 1.

$$\delta(d_{min},d_{max}) = \frac{PN(d_{min},d_{max}) - N(d_{min},d_{max})}{N(d_{min},d_{max})} \qquad (3)$$

Two useful monotonically increasing functions of x, called the *forward* and *backward distributions*, are defined as follows:

$$f(x) = N[0,x] / N[0,\infty] \qquad (4)$$

and

$$b(x) = (N[0,\infty] - N[x,\infty]) / N[0,\infty] \qquad (5)$$

The projected number of hits can be written in terms of these two functions:

$$PN[d_{min},d_{max}] = (f(d_{max}) - b(d_{min})) / N[0,\infty] \qquad (6)$$

Thus, the projected number of hits for any set of query composed of a single distance constraint between two features, a *dyad* query, may be computed from a table of forward and backward distributions for these feature pairs.

Catalyst v. 2.1 [9] was used for database searching and database construction. In that version of Catalyst, conformational flexibility is treated by representing each molecule with multiple conformers, up to some specified maximum number of conformers. The set of conformers used to describe each molecule is chosen to efficiently describe conformational space by a careful selection of conformers [14]. No adjustments of the conformer are performed on-the-fly in assessing whether a conformer satisfies the search query; a simple strict match is made between each conformer and the query.

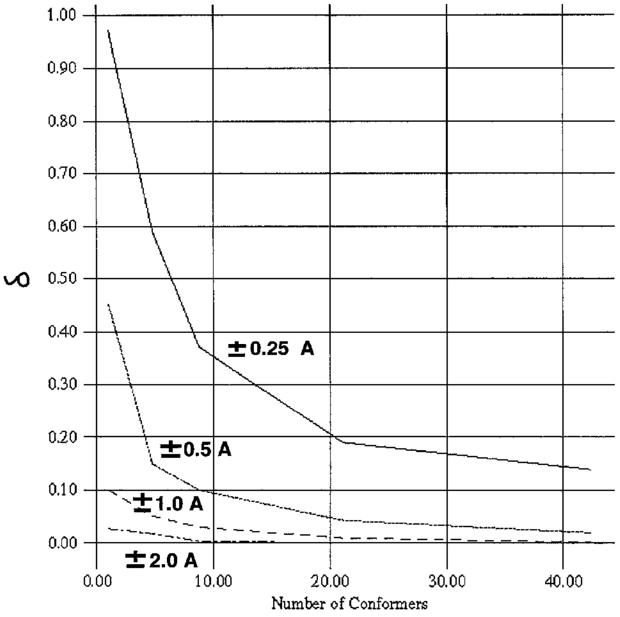

Fig. 1. NCI database: δ as a function of the number of conformers and the precision of the query.

A single-conformer and a set of multiconformer databases containing up to 5, 10, 30, and 100 conformers (hereafter referred to by the suffixes X1, X5, X10, X30, and X100) were constructed. Two different sets of molecules were used as input: the NCI database of 99 000

TABLE 3

OBSERVED AND PROJECTED NUMBERS OF HITS USING THE QUERY OF BASIC NITROGEN 6 ± η Å FROM A CARBOXYL AGAINST THE DATABASE OF DI- AND TRIPEPTIDES CONSTRUCTED WITH VARYING NUMBERS OF CONFORMERS/ MOLECULE

| η (Å) | Database | N[6 − η, 6 + η] | N[0, 6 + η] | N[6 − η, ∞] | PN[6 − η, 6 + η] | δ[6 − η, 6 + η] |
|---|---|---|---|---|---|---|
| 2.0 | pepX1 | 4125 | 4190 | 8400 | 4190 | 0.015 |
| 1.0 | pepX1 | 1901 | 2530 | 8386 | 2516 | 0.32 |
| 0.5 | pepX1 | 1135 | 2277 | 8374 | 2251 | 0.98 |
| 0.25 | pepX1 | 660 | 2170 | 8352 | 2122 | 2.22 |
| 2.0 | pepX5 | 8387 | 8392 | 8400 | 8392 | 0.0006 |
| 1.0 | pepX5 | 7573 | 7991 | 8400 | 7991 | 0.055 |
| 0.5 | pepX5 | 5169 | 7066 | 8400 | 7066 | 0.37 |
| 0.25 | pepX5 | 3113 | 6538 | 8388 | 6526 | 1.10 |
| 2.0 | pepX10 | 8399 | 8400 | 8400 | 8400 | 0.0001 |
| 1.0 | pepX10 | 8273 | 8372 | 8400 | 8372 | 0.012 |
| 0.5 | pepX10 | 6822 | 8102 | 8400 | 8102 | 0.19 |
| 0.25 | pepX10 | 4625 | 7847 | 8391 | 7838 | 0.70 |
| 2.0 | pepX30 | 8400 | 8400 | 8400 | 8400 | 0 |
| 1.0 | pepX30 | 8400 | 8400 | 8400 | 8400 | 0 |
| 0.5 | pepX30 | 8121 | 8393 | 8400 | 8393 | 0.033 |
| 0.25 | pepX30 | 6891 | 8384 | 8395 | 8379 | 0.22 |
| 2.0 | pepX100 | 8400 | 8400 | 8400 | 8400 | 0 |
| 1.0 | pepX100 | 8400 | 8400 | 8400 | 8400 | 0 |
| 0.5 | pepX100 | 8341 | 8398 | 8400 | 8398 | 0.007 |
| 0.25 | pepX100 | 7858 | 8398 | 8396 | 8394 | 0.07 |

N[0,∞] is 8400 in each case. The data are plotted in Fig. 2.

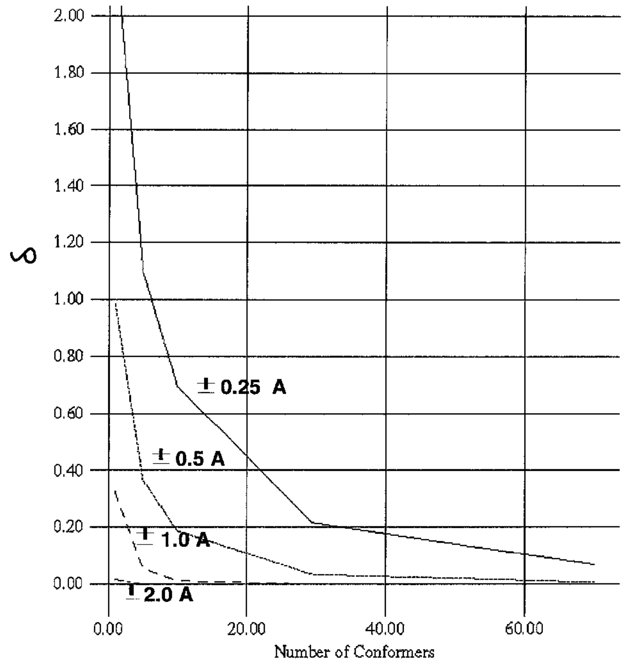

Fig. 2. Database of di- and tripeptides: δ as a function of the number of conformers and the precision of the query.

molecules [15] and an artificially constructed database of the 8400 naturally occurring di- and tripeptides. Table 1 shows that the actual average number of conformers in each database was less than the maximum specified at construction, e.g. the average number of conformers for each molecule in NCIX10 is 8.77. Not surprisingly, the average number of conformers per molecule for the database of drug-like molecules is consistently smaller than that for the database of peptides since drug-like molecules are generally less conformationally flexible than peptides.

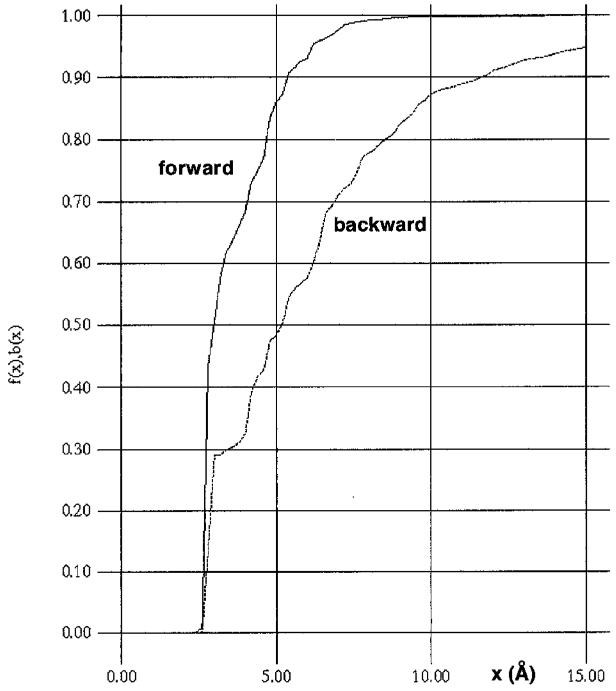

Fig. 3. NCI database: forward and backward distributions.

Database searches were performed on each database for a set of queries, each consisting of a distance constraint connecting the center of a carboxyl group and a basic nitrogen. Queries were composed for distance ranges $[6-\eta, 6+\eta]$, where $\eta = 2$, 1, 0.5, and 0.25 Å. The number of hits was tallied for each of the queries, $N[6-\eta, 6+\eta]$, against each of these databases. Additional queries were composed, and searches were performed to allow the computation of the projected number of hits, i.e., for each $N[d_{min},d_{max}]$, searches were performed to determine $N[0,d_{max}]$, $N[d_{min},\infty]$, and $N[0,\infty]$. Relative differences (δ, see Eq. 3) were then computed based on these projected numbers of hits.

All of the search queries used in this work contained only two features (topological constraints) connected by one distance constraint. One expects to be able to compute the projected number of hits for three-feature queries via the equation

$$\mathrm{PN(A\text{–}B\text{–}C)} = \frac{\mathrm{N(A\text{–}B)}}{\mathrm{N(AB)}}\,\frac{\mathrm{N(A\text{–}C)}}{\mathrm{N(AC)}}\,\frac{\mathrm{N(B\text{–}C)}}{\mathrm{N(BC)}}\,\mathrm{N(ABC)} \qquad (7)$$

where N(AB) is the number of hits retrieved with no distance constraint and N(A–B) the number of hits retrieved with a distance constraint. However, this relation usually differs from N(A–B–C) by more than 100%, presumably due to a lack of statistical independence between the arrangement of different features. Another possible route to solving this problem may be the generalization of the arguments presented in the Appendix to higher dimension queries.

## Results and Discussion

Table 2 shows the tallied numbers of hits for the different versions of the NCI database; Fig. 1 depicts these data graphically, with a family of curves of δ versus the average number of conformers, one curve per precision. Table 3 and Fig. 2 show the corresponding data for the database of peptides.

A number of conclusions can be drawn from these data:

(1) The δ's remain below 0.10 for single-conformer databases when the queries are imprecise (± 2 Å). This suggests that 3D database searching of single-conformer databases is acceptable, provided the query is suitably imprecise. This conclusion is similar to the one made by Güner et al. [16].

(2) The δ's improve steadily as we have steadily improved the treatment of conformational flexibility. This conclusion is likely to hold true for those methods which

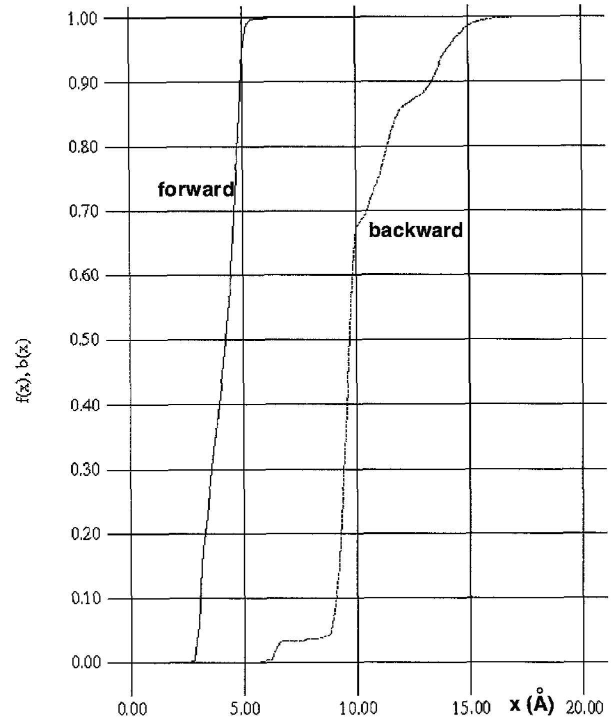

Fig. 4. Database of di- and tripeptides: forward and backward distributions.

treat conformational flexibility on-the-fly; similar studies need to be made for those techniques.

(3) Even at large numbers of conformers per molecule, we may regard our queries as precise only to ±0.5 Å, if we wish to keep the magnitude of δ to less than 0.03.

(4) A complementarity (or trade-off) exists between how well we treat conformational flexibility and how precise we may regard our queries. Graphs such as in Figs. 1 and 2 can be used to show, for a given database, a given target δ and a given degree of treatment of conformational flexibility, what is the maximum precision we may ascribe to our queries.

Figure 3 shows the ‘forward distribution’ and ‘backward distribution’ for this class of query applied to the NCI database; Fig. 4 shows the same for the peptide database. As expected, these distributions are monotonic (failure to demonstrate monotonicity is indicative of either software bugs or weaknesses in the algorithm for treating conformational flexibility).

## Conclusions

An inequality has been presented, which leads to the definition of a quantity δ which may be used to quantitate the trade-offs between the precision of a 3D database query and the demands on the quality of the treatment of conformational flexibility. We have shown that δ is typically quite small when the queries are imprecise even when conformational flexibility is not treated at all (single-conformer database), but that with precise queries it is important to treat conformational flexibility well to ensure small δ's. These results have been shown only with multiple-conformer databases, but they should generally be true for all methods of treating conformational flexibility in 3D database searching. Also, this inequality holds only for database queries containing a single distance constraint; how it may generalize to higher dimension queries is still unclear.

## Acknowledgements

The author is grateful to G.M. Maggiora for his help in clarifying the arguments in the Appendix.

## References

1 Gund, P., Wipke, T. and Langridge, R., Proc. Int. Conf. Comput. Chem. Res. Ed., 3 (1974) 5.

2 Willett, P., 3D Chemical Structure Handling, Wiley, New York, NY, U.S.A., 1991, and references cited therein.

3 Sheridan, R.P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 29 (1989) 255.

4 a. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
b. DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 29 (1986) 2149.

5 VanDrie, J.H., Weininger, D. and Martin, Y.C., J. Comput Aided Mol. Design, 3 (1989) 255.

6 Christie, B.D., Henry, D.R., Güner, O.F. and Moock, T.E., Online Inf., 90 (1990) 137.

7 Murrall, N.W. and Davies, E.K., J. Chem. Inf. Comput. Sci., 30 (1990) 312.

8 Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.

9 VanDrie, J.H., Berezin, S. and Ku, S.-L., Abstracts for the ACS National Meeting, Spring, 1992, CINF 023. No more detailed descriptions of the Catalyst 3D database searching software have been published. The software is available from Molecular Simulations Inc., San Diego, CA, U.S.A.

10 a. Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., In Roberts, S.M. (Ed.) Molecular Recognition: Chemical and Biological Problems, Vol. 78, Royal Society of Chemistry, London, U.K., 1989, pp. 182–192.
b. Lauri, G. and Bartlett, P.A., J. Comput.-Aided Mol. Design, 8 (1994) 51.

11 CONCORD, Rusinko, A., Skell, J., Balducci, R., McGarity, M.C. and Pearlman, R.S.; A description of this methodology may be found in the Ph.D. Thesis of Rusinko III, A. (1988), University of Texas, Austin, TX, U.S.A., available from University Microfilms, Ann Arbor, MI, U.S.A.

12 Moock, T.E., Henry, D.R., Ozkabak, A.G. and Alamgir, M., J. Chem. Inf. Comput. Sci., 34 (1994) 184.

13 a. Kearsley, S.K., Underwood, D.J., Sheridan, R.P. and Miller, M.D., J. Comput.-Aided Mol. Design, 8 (1994) 565.
b. Miller, M.D., Kearsley, S.K., Underwood, D.J. and Sheridan, R.P., J. Comput.-Aided Mol. Design, 8 (1994) 153.

14 Smellie, A., Teig, S. and Towbin, P., J. Comput. Chem., 16 (1995) 171.

15 NCI database, available from the National Cancer Institute, Frederick, MD, U.S.A.

16 Güner, O.F., Henry, D.R. and Pearlman, R.S., J. Chem. Inf. Comput. Sci., 32 (1992) 101.

# Appendix

## Derivation of fundamental inequality

We will develop the fundamental inequality stepwise, starting with simple counting arguments, based on the notion of the *characteristic function* of a molecule for a given 'dyad' query, a query constraining two features to lie within a given distance range.

### *The characteristic function of a molecule*

The characteristic function, $\chi$, is the number of hits that is returned for a database search against a database consisting of just that one molecule, where the distance constraint of the dyad query is from x to $x+\eta$, where $\eta$ is a small number and x ranges from 0 to $\infty$. Hence, it is a function which takes on values 0 or 1.

The characteristic function for alanine for a dyad query of any N to any O is shown in Fig. 5. For values of $x<2.5$ Å, the characteristic function is zero: the molecule cannot adopt an energetically reasonable conformation which holds any O that distance or less from the N. For values of x in the range 2.5–3.7, $\chi$ is 1; here the query may map to either of the oxygens and the nitrogen, as the molecule adopts either a strained conformation which holds the termini relatively closely, or one which is extended, or any conformation between these extremes. At $x>3.7$, $\chi$ again is zero, as the query is unable to map to a fully extended alanine with that distance constraint.

The characteristic function of the dipeptide serine-lysine for an N-O dyad is shown in Fig. 6. The smallest value of x for which $\chi$ is nonzero, 2.5 Å, corresponds to the mapping of the terminal N and the serine side chain. Intermediate values of x for which $\chi$ is nonzero correspond to many possible mappings (e.g. side-chain–side-chain, side-chain–amide bond, side-chain–terminus). The largest values for which $\chi$ is nonzero correspond to mappings of the query to the termini, 12.0 Å.

For a given characteristic function, the smallest value of x for which $\chi$ is nonzero will be referred to as the 'turn-on point'; the largest value of x for which $\chi$ is nonzero will be referred to as the 'turn-off point'. In general, a characteristic function will have regions in which it is zero between the turn-on and turn-off points (multimodal distribution). In developing our inequality, we will begin by assuming that all characteristic functions do not have this property, i.e. they are constant between their turn-on and turn-off points (unimodal). At the end of this derivation, we will relax this 'unimodality' assumption.

Figure 7 shows the unimodal characteristic functions of three different molecules for a given dyad. Molecule A turns on at 3 Å and turns off at 10 Å; molecule B turns on at 4 Å and turns off at 9 Å; molecule C turns on at 5 Å and turns off at 12 Å. Given a database consisting only of these three molecules, the number of hits which would be returned by a dyad with a distance range of 5–7 Å would be 3, i.e. $N[5,7]=3$; all three molecules may adopt conformations which place that pair of features some-

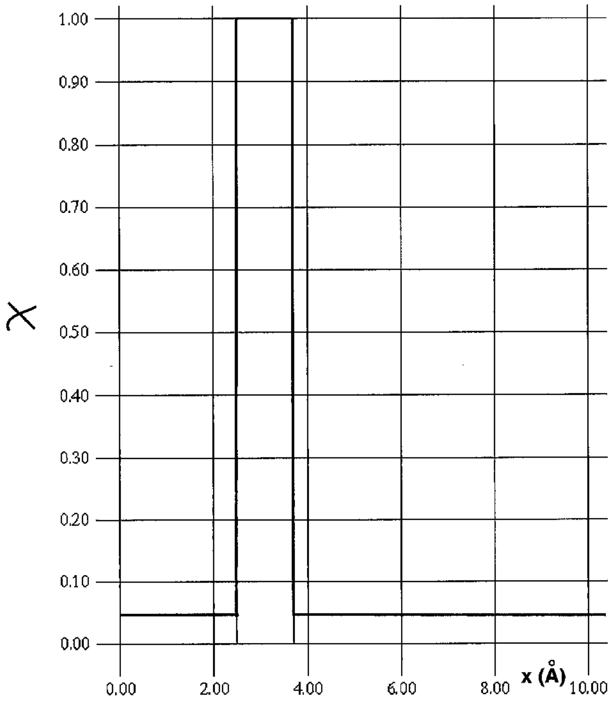

Fig. 5. Characteristic function for alanine, for a dyad query of any nitrogen to any oxygen.

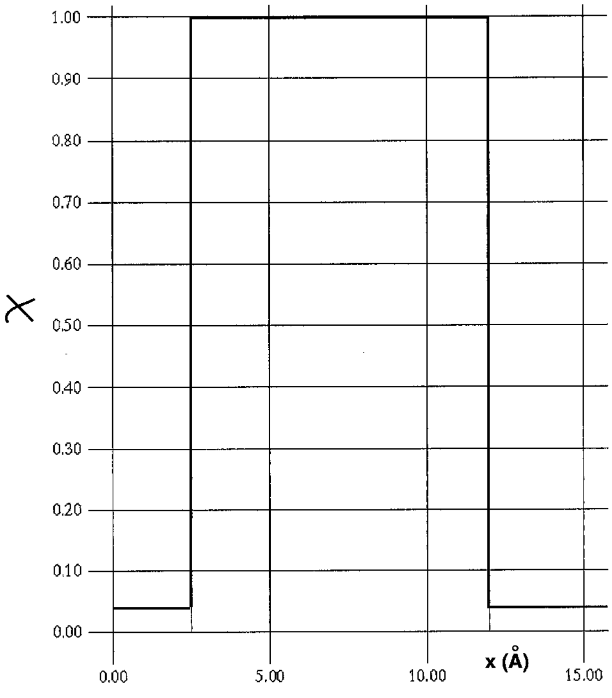

Fig. 6. Characteristic function for the serine-lysine dipeptide, for a dyad query of any nitrogen to any oxygen.

where in the range [5,7]. By inspection of these characteristic functions, we see that $N[1,3.5]=1$ (only one molecule has a nonzero $\chi$ in the range [1,3.5], molecule A) and $N[9.5,15]=2$ (molecules A and C both have nonzero $\chi$'s in that range).

*Equality 1* Assume that all molecules in our database have unimodal characteristic functions, and denote by $N[d_{min},d_{max}]$ the number of hits for a given dyad when the distance between the features is constrained to be in the range $[d_{min},d_{max}]$. It is apparent by simple counting that

$N[d_{min},d_{max}]$ = the number of molecules whose $\chi$'s have turn-on points $\leq d_{max}$
− the number of molecules whose $\chi$'s have turn-off points $\leq d_{min}$

We can apply this equality to the previous example of a three-molecule database. For a range [5,7], the number of molecules whose characteristic functions have turned on by 7 Å is 3, and the number of molecules whose $\chi$'s have turned off by 5 Å is 0; hence $N[5,7]=3-0=3$, as noted earlier. For the range [1,3.5], the number of molecules whose $\chi$'s have turned on by 3.5 is 1 (molecule A), while the number that have turned off by 1 is 0; hence $N[1,3.5] = 1-0=1$. For the range [9.5,15], the number which have turned on by 15 is 3, while the number which have turned off by 9.5 is 1 (molecule B); hence $N[9.5,15]=3-1=2$.

*Definitions leading to a second equality derived from the first*

Let us partition the x-axis into a series of arbitrarily sized segments (e.g. partition $P_1=[0,1)$, $P_2=[1,2)$, $P_3=[2,3)$, etc.). Denote the starting and ending limits of each partition by $P_i^{START}$ and $P_i^{END}$, respectively. Denote by $F_i$ the number of molecules in the database whose $\chi$'s have turn-on points in the partition $P_i$. Denote by $B_i$ the number of molecules in the database whose $\chi$'s have turn-off points in the partition $P_i$.

*Equality 2* Continue to assume that all molecules have unimodal $\chi$'s. We may now use the $F_i$'s and $B_i$'s to write an explicit expression for the number of molecules whose $\chi$'s have turn-on points $\leq d_{max}$, and the number of molecules whose $\chi$'s have turn-off points $\leq d_{min}$. It is clear from simple counting that

the number of molecules
whose $\chi$'s have turn-on points $\leq d_{max} = \Sigma F_i$

where the summation is taken over all partition segments from 0 to $d_{max}$. Similarly,

the number of molecules
whose $\chi$'s have turn-off points $\leq d_{min} = \Sigma B_i$

where the summation is taken over all partition segments from 0 to $d_{min}$. Hence, equality 1 becomes

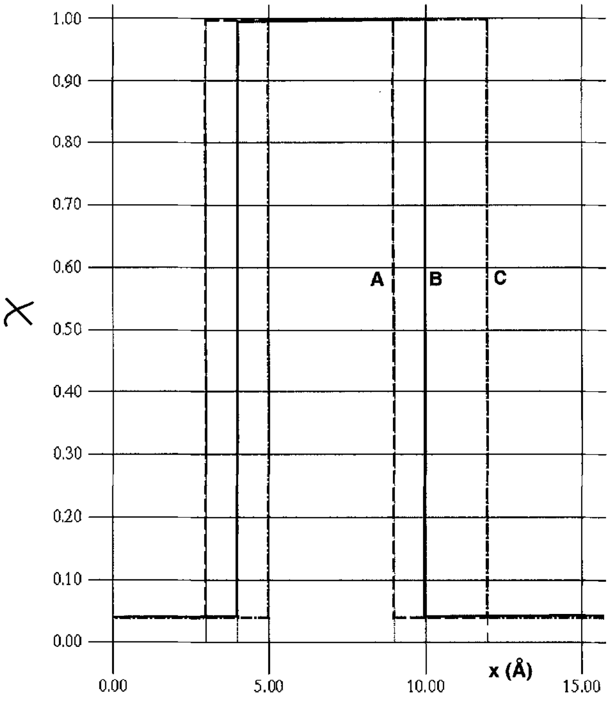

Fig. 7. Characteristic functions of hypothetical molecules A, B, and C for a hypothetical dyad query.

$$N[d_{min},d_{max}] = \Sigma_{[0,d_{min}]} F_i - \Sigma_{[0,d_{max}]} B_i$$

*Equality 3* Continue with the unimodality assumption. We can convert the right-hand side of equality 2 into a set of terms which one can measure via database searching. Note that one can measure $F_i$ and $B_i$ directly via database searching: $F_i$ is the difference between the number of database hits returned using the distance range $[0,P_i^{END}]$ and the number of hits returned using the range $[0,P_i^{START}]$, and $B_i$ is the difference between the number of hits returned using the range $[P_i^{START},\infty]$ and the number of hits returned using the range $[P_i^{END},\infty]$. By cancellation of terms, we then see that

$\Sigma_{[0,d_{max}]} F_i$ = number of database hits returned using the range $[0,d_{max}]$
− number of database hits returned using the range [0,0]

$\Sigma_{[0,d_{min}]} B_i$ = number of database hits returned using the range $[0,\infty]$
− number of database hits returned using the range $[d_{min},\infty]$

The number of hits in the range [0,0] must be zero; hence, equality 2 may be rewritten as

$N[d_{min},d_{max}]$ = number of hits in the range $[0,d_{max}]$
+ number of hits in the range $[d_{min},\infty]$
− number of hits in the range $[0,\infty]$

*Relaxation of the unimodality assumption*

Up to this point, we have assumed that all characteristic functions are unimodal, i.e. they are constantly 1 between their turn-on point and their turn-off point. We must now inspect the implications of this assumption. Refer now to the right-hand side of equality 3 as $PN[d_{min},d_{max}]$, the projected number of hits in the range $[d_{min},d_{max}]$, assuming all $\chi$'s are unimodal. If a molecule has a $\chi$ which is actually zero in the range $[d_{min},d_{max}]$, but has a turn-on point less than $d_{max}$ and a turn-off point less than $d_{min}$, we scc that $PN[d_{min},d_{max}]$ incorrectly counts that molecule, while it would not appear as a hit in a database search. Hence, by assuming all $\chi$'s are unimodal, $PN[d_{min},d_{max}]$ consistently overcounts the actual number of hits returned from a database search with a dyad whose distance range is $d_{min}$ to $d_{max}$. In other words,

$PN[d_{min},d_{max}] \geq$ number of hits in the range $[d_{min},d_{max}]$

or

number of hits in the range $[0,d_{max}]$
+ number of hits in the range $[d_{min},\infty]$
− number of hits in the range $[0,\infty]$
≥ number of hits in the range $[d_{min},d_{max}]$

which is our desired inequality.

*Impact of deficiencies in the treatment of conformational flexibility in molecules*

Up to this point, we have implicitly assumed that the treatment of conformational flexibility in the 3D database search is perfect, which can never be true. We now consider the impact of deficiencies in such methods in the inequality. The predominant source of deficiencies is an inadequate sampling of conformational space. Admittedly, it is possible for an algorithm for conformationally flexible 3D database searching to return a molecule as a hit when it should not be (e.g. by placing two atoms a distance apart that is significantly smaller than the sum of their van der Waals radii), but simple energy tests generally filter out such errors. The impact of an inadequate sampling of conformational space on the characteristic function of a molecule would be to introduce nodes (regions where $\chi$ is zero) where $\chi$ actually should be 1, i.e. a molecule should hit a given query but does not. Hence, in a realistic setting, the inequality should still hold, but the sources of the non-unimodality may come from both the molecules actually possessing multimodal $\chi$'s (in the limit of perfect conformational analysis) or from inadequate sampling of the conformational space of the molecule.