communications in mathematical
match
and in computer chemistry

no. 45, March 2002

MATCDY (45) 153 - 172 (2002)

ISSN 0340 - 6253

How Many Organic Compounds Are Graph-Theoretically Nonplanar?

Christoph Rücker* and Markus Meringer
Department of Mathematics, Universität Bayreuth,
D-95440 Bayreuth, Germany
e-mail: Christoph.Ruecker@uni-bayreuth.de

Abstract. Most graphs and most 4-graphs are nonplanar, whereas most compounds of Organic Chemistry are gt-planar. This potentially useful fundamental difference between graphs and compounds was empirically obtained by testing for gt-planarity representative graphs and all compounds in the Beilstein file. The Beilstein search did uncover compounds whose gt-nonplanarity was hitherto unknown. gt-Nonplanar peptides/proteins were retrieved from the CAS Registry file.

INTRODUCTION

In computer-assisted structure elucidation, e.g. of natural products, a graph generator is used to construct all possible candidate structures (colored multigraphs) for a given molecular formula under a set of restrictions derived from various spectroscopic measurements. Often one is confronted with a very large number of candidates, therefore further restrictions are urgently needed.

For an unknown containing more than three cycles (and these are the more difficult problems), as a rule many of the candidate structures are gt-nonplanar, whereas many chemists feel that gt-nonplanar compounds are rare.[1] The term *gt-nonplanar* (short for *graph-theoretically nonplanar*) means that the constitutional formula (structural formula) of such a compound cannot be drawn in a plane without at least two bonds crossing. Most known

compounds, even geometrically decidedly nonplanar ones such as helicenes, fullerenes, and most other synthesized polycyclic structures, correspond to planar graphs, i.e. their structural formulae (severely distorted, if necessary) can be drawn on a sheet of paper without two bonds crossing.

For example, in a (simulated) structure elucidation of heptacyclic strychnine ($C_{21}H_{22}N_2O_2$), on a particular stage the computer program proposed five structures, four of them gt-nonplanar and one planar, the latter was the correct structure.[2] Thus it might be a good idea to optionally exclude gt-nonplanar candidate structures from consideration on an early stage, or not even to generate them. Obviously a considerable difference in the frequency of nonplanars between graphs on the one hand and compounds on the other is a prerequisite for this idea to be useful.

Actually, up to now nobody knows how many compounds, and which ones, correspond to nonplanar graphs, i.e. nobody has a complete list of such structures. Obviously, the number of such compounds depends on what kind of interatomic interaction is represented by a line. As usual in Organic Chemistry, we here restrict ourselves to consider covalent bonds only.[3]

The first compound synthesized purposefully to exhibit this very exclusive property was published in 1981.[4,5] Ironically, the first synthesized gt-nonplanar organic compounds (**1**, two diastereomers) were obtained a few months earlier, but the investigators were not aware of gt-nonplanarity.[6] Since then several gt-nonplanar compounds were synthesized episodically, with the respective authors sometimes being aware of this fact,[7-9] sometimes not.[10-13] Not surprisingly, in a recent literature survey on gt-nonplanar molecules[14] several pertinent compounds are missing, partially because nobody had recognized them as such.[11-13,15] The gt-nonplanar compounds shown in Figure 1 are representative of these. It is therefore perfectly possible that some gt-nonplanar compounds still went unnoticed. In fact, the two gt-nonplanar organic compounds to appear first in the literature (natural products, highly disulfide-bridged proteins isolated in 1974[15]) were not recognized as gt-nonplanar until many years later.[16] Interestingly, all gt-nonplanar natural products known to us (except the very special case of diamond) belong to this class. This is equivalent to the statement that not a single natural product known to us at present (except diamond and some peptides/proteins) is gt-nonplanar.

In Appendix I all non-peptidic gt-nonplanar compounds of which we know at present are listed.

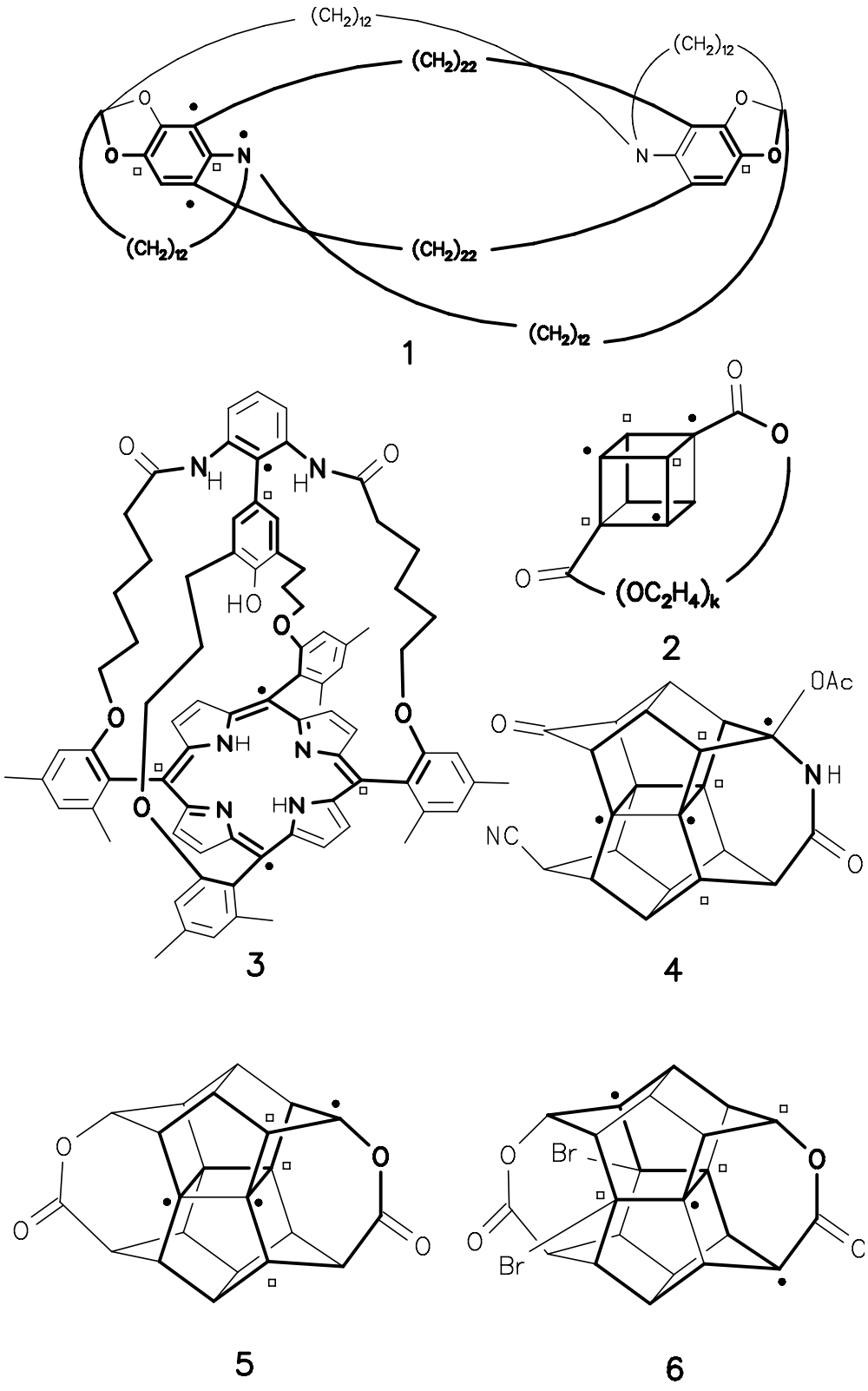

Figure 1. Some gt-nonplanar compounds not recognized as such earlier. All are homeomorphic to $K_{3,3}$, black dots and open squares are used to indicate this fact. Atoms and bonds included in the homeomorphic $K_{3,3}$ are drawn in heavy lines.

In the present work, therefore, we undertook to test for gt-(non)planarity representative graphs and to do a computer-assisted search for gt-nonplanars among the known molecular structures of Organic Chemistry, in order to provide an empirical basis for further studies.

METHOD

Mathematically, it is quite clear which graphs are nonplanar: As proven by Polish mathematician K. Kuratowski, for a graph to be nonplanar it is both necessary and sufficient to contain as a subgraph a graph homeomorphic to one of the two graphs known as K_5 and $K_{3,3}$ (Figure 2).[17,18] Since $K_{3,3}$ is tetracyclic (K_5 is hexacyclic), a gt-nonplanar structure contains at least one at least tetracyclic ring system. To prove the gt-nonplanarity of a particular compound it is therefore sufficient to demonstrate the presence of a K_5 or $K_{3,3}$ homeomorph in its structure. All compounds in Figure 1 are $K_{3,3}$ homeomorphs as indicated by heavy lines between atoms marked with black dots and open squares. The reader may convince himself that from each black dot there are disjoint paths leading to the three open squares.

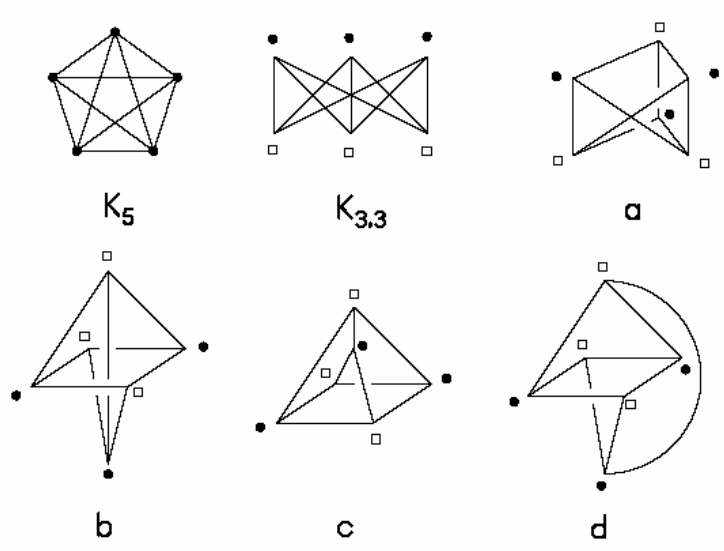

Figure 2. Nonplanar graphs K_5 and $K_{3,3}$ and a few 3D realizations of the latter. Black dots and open squares indicate two sets of vertices, each vertex is connected via disjoint paths to each vertex in the other set. From **a-d** a recipe is derived to obtain a gt-nonplanar compound homeomorphic to $K_{3,3}$: Doubly and orthogonally bridge a central cycle, and interconnect the bridges.

The number ν of atoms forming the homeomorphic K_5 or $K_{3,3}$ is a convenient measure of the size of a gt-nonplanar system. Thus, **1** has ν = 84, **2**, *k* = 4, 5, 6 have ν = 23, 26, 29, respectively, **3** has ν = 56, **4** and **5** have ν = 15, and **6** has ν = 18. Of the gt-nonplanar compounds known earlier, Walba's 3-rung and 4-rung Möbius strips[7] have ν = 60 and 80, respectively, Otsubo's triple-layer naphthalenophanes[10] have ν = 30 and 34, and Siegel's Kuratowski cyclophane[9] has ν = 70. The gt-nonplanar proteins isolated in 1974 (64 and 66 amino acids, respectively[15]) have ν = 170 and 176. The Simmons/Paquette compound[4,5] and most of Kuck's K_5 homeomorphs[8] have ν = 17.

Interestingly, graph-theoretical planarity tests available as algorithms or computer programs do not rely on the Kuratowski criterion. Rather, such tests literally try to embed the structure in a plane without crossing of edges, a procedure more efficient than finding subgraphs of unknown structure.[19] An implementation of the Lempel-Even-Cederbaum planarity test[20] written in C++ (a part of the excellent Graph Template Library (GTL)[21]) was provided by Professor F. J. Brandenburg of the University of Passau, Germany.

Testing the test. In order to convince ourselves of the correctness and efficiency of this particular implementation, we used it to test for gt-(non)planarity several samples of graphs for which the correct results are known:

- The graphs corresponding to several gt-nonplanar compounds and to their gt-planar analogs, such as the Simmons-Paquette compound and its planar precursor,[4,5] the Schill macropolycycle **1**,[6] the Walba 3-rung Möbius strip compound and its prism analog,[7] the Otsubo triple-layer naphthalenophane,[10] a Moriarty bridged cubane **2**,[13] a Prinzbach bridged pagodane **4**.[11]
- The five strychnine candidate structures mentioned above.[2]
- The cubic graphs of 4, 6, 8, …, 26 vertices, and the quartic graphs of 5, 6, …, 18 vertices,[22] generated using program GENREG.[23] The numbers of planar graphs found among these[24a] agree with those published,[24b] as far as such numbers are available.

The planarity test not only gave correct results in all these cases, it also is rather fast: To process the sample of 510489 cubic graphs of 20 vertices (no more than 24809 planar graphs among them) 411 sec were required on a 800 MHz Pentium III PC under Windows NT.

RESULTS AND DISCUSSION

Testing graphs. Since gt-(non)planarity of an organic compound is independent of the presence of multiple bonds or heteroatoms, for planarity considerations it suffices to deal with simple graphs. First we generated (MOLGEN 4.0[25]) and tested for gt-(non)planarity all connected simple undirected graphs of $n = 9$ vertices, in classes of constant edge numbers m, that is also constant numbers of cycles c ($c = m - n + 1$). Among these all graphs of $c \leq 3$ were found to be planar, all those of $c \geq 14$ were found to be nonplanar,[26] and for $4 \leq c \leq 13$ the percentage of planar graphs among all such graphs monotonically decreases from 99.09 to 0.23%. Results are shown in Table 1.

Table 1. Numbers of graphs and planar graphs of constant vertex number n dependent on edge number m or cycle number c for two classes of simple graphs.

simple graphs of $n = 9$

m	c	#	planar	% planar
8	0	47	47	100
9	1	240	240	100
10	2	797	797	100
11	3	2075	2075	100
12	4	4495	4454	99.09
13	5	8404	8053	95.82
14	6	13855	11990	86.54
15	7	20303	14379	70.82
16	8	26631	13380	50.24
17	9	31400	9464	30.14
18	10	33366	4844	14.52
19	11	31996	1734	5.42
20	12	27764	378	1.36
21	13	21817	50	0.23
22	14	15558	0	0
23	15	10096	0	0
24	16	5984	0	0
25	17	3247	0	0
26	18	1635	0	0
27	19	770	0	0
28	20	344	0	0
29	21	148	0	0
30	22	63	0	0
31	23	25	0	0
32	24	11	0	0
33	25	5	0	0
34	26	2	0	0
35	27	1	0	0
36	28	1	0	0
total		261080	71885	27.53

simple 4-graphs of $n = 12$

m	c	#	planar	% planar
11	0	355	355	100
12	1	3232	3232	100
13	2	16977	16977	100
14	3	64720	64720	100
15	4	192945	190632	98.80
16	5	460699	428737	93.06
17	6	880737	696250	79.05
18	7	1327240	781796	58.90
19	8	1530906	581150	37.96
20	9	1290182	271833	21.07
21	10	737981	73604	9.97
22	11	252720	10152	4.02
23	12	40399	551	1.36
24	13	1544	13	0.84
total		6800637	3120002	45.88

For the set of organic compounds the set of all simple 4-graphs (graphs containing no vertex of degree higher than four) is a closer model than the set of all simple graphs. We therefore performed a corresponding test on all simple 4-graphs of $n = 12$ vertices, in classes of constant m or c. Again, all graphs of $c \leq 3$ were found planar, while for the classes of $4 \leq c \leq 13$ the percentage of planar graphs monotonically decreases from 98.80 to 0.84% (Table 1).

Second we tested the connected simple undirected graphs of $n = 1, 2, 3, \ldots, 10$ vertices without sorting by m, see Table 2. The numbers of planar graphs found agree with published data.[24b] With increasing n the percentage of planar graphs monotonically decreases. Beginning with $n = 9$ the majority of graphs is nonplanar, and since this trend will undoubtedly continue for higher n, we can conclude that most graphs are nonplanar.

The corresponding test was performed for all 4-graphs of $n = 1, 2, 3, \ldots, 12$ (Table 2). Again, for increasing n the percentage of planar 4-graphs monotonically decreases. From $n = 12$ onwards, the majority of 4-graphs is nonplanar, and thus most 4-graphs are nonplanar as well.

Table 2. Numbers of graphs and planar graphs, and of 4-graphs and planar 4-graphs of n vertices.

	simple graphs			simple 4-graphs			
n	#	planar	% planar	n	#	planar	% planar
1	1	1	100	1	1	1	100
2	1	1	100	2	1	1	100
3	2	2	100	3	2	2	100
4	6	6	100	4	6	6	100
5	21	20	95.24	5	21	20	95.24
6	112	99	88.39	6	78	74	94.87
7	853	646	75.73	7	353	321	90.93
8	11117	5974	53.74	8	1929	1663	86.21
9	261080	71885	27.53	9	12207	9616	78.77
10	11716571	1052805	8.99	10	89402	61840	69.17
				11	739335	427135	57.77
				12	6800637	3120002	45.88

In view of the gt-nonplanar proteins mentioned above we also generated and tested for gt-(non)planarity all disulfide bridge patterns of linear proteins with $c = 1, 2, 3, \ldots, 9$ disulfide bridges, which correspond to special c-cyclic graphs. Results are given in Table 3, to be compared with an earlier but less complete table.[27] With increasing number of disulfide

bridges, the percentage of gt-planar patterns monotonically decreases. Starting with eight disulfide bridges, the majority of patterns is nonplanar, so that we can conclude that most disulfide bridge patterns are nonplanar.

Table 3. Numbers of linear protein disulfide bridge patterns and of planar such patterns for *c* disulfide bridges.

c	#	planar patterns	% planar patterns
1	1	1	100
2	3	3	100
3	15	15	100
4	105	101	96.19
5	945	815	86.24
6	10395	7373	70.93
7	135135	71943	53.24
8	2027025	739847	36.50
9	34459425	7903741	22.94

Testing organic compounds. Information on gt-(non)planarity of organic compounds is implicitely present in the molecular structures deposited in the huge chemistry databases CAS Registry file and MDL Beilstein file. However, there are several technical, legal and financial obstacles against retrieval of this information.

Fortunately, instead of testing all compounds it would suffice to test all at least tetracyclic ring systems. Far fewer ring systems than compounds exist, and even fewer tetra- or highercyclic ring systems.[28] Information on which compounds contain which ring systems is available in the online versions of the files accessible via STN. Unfortunately, while the ring systems are available (at high cost) from the Registry file, they are not obtained in computer readable format, and in Beilstein Crossfire (the version of Beilstein available in many academic institutions) the ring systems are not searchable, while downloading large parts of the database (the compounds) is prohibited. We therefore were not able to perform the planarity test of organic compounds ourselves. Instead, we asked the database producers for help, and Dr. D. Hounshell of MDL agreed to run the planarity test on all Beilstein compounds at the MDL facilities.

Testing peptides/proteins. Knowing that gt-nonplanar peptides/proteins do exist,[15,16] we were interested in a list of such species as complete as possible. Unfortunately, peptides and proteins are not covered by the Beilstein file. They are included in the CAS Registry file, but as amino acid sequences only rather than as structures. Disulfide bridges are registered as modifications of the sequence. Even worse, this registering practice seems to have become effective not before the 1990s. Despite these serious limitations, in order to obtain proteins with at least 4 known disulfide bridges we searched the Registry file for proteins with at least eight cysteine residues and the entry "disulfide" in the note field. This resulted in 955 proteins, most of them published since the mid-1990s. These were manually checked for their number of peptide chains, number and pattern of disulfide bridges, gt-planarity, and origin (natural or man-made). From the 955 proteins the following had to be eliminated: i) obviously erroneous entries having the same cysteine residue engaged in two disulfide bridges, ii) multichain proteins, iii) linear peptides/proteins with fewer than four disulfide bridges assigned, iv) circular peptides with less than three disulfide bridges.[29] Being mostly interested in the planar/nonplanar distribution among natural peptides/proteins, we further eliminated all man-made species, which finally resulted in 356 natural peptides/proteins. The oldest nonplanar protein among these was published in 1992. We are aware that this small database probably reflects current trends in peptide/protein research more than the actual planar/nonplanar distribution in natural proteins. Nevertheless the results shown in Table 4 may be of some interest. They at least reveal that in contrast to what would be expected based simply on statistics (Table 3), the percentage of known gt-nonplanar proteins among all known proteins of a specific number of disulfides does not increase with increasing number of disulfide bridges. On the contrary, not a single confirmed case of a gt-nonplanar protein with more than six disulfide bridges is known to us.[30]

38 peptides/proteins out of the 356 do not appear in Table 4. These are circular peptides of three disulfide bridges, and their majority (30) is gt-nonplanar, see Class 5 below.

The gt-nonplanar peptides/proteins from the Registry file search are listed in Appendix II, first part (including man-made ones). Due to the limitations in CAS' registering policy, several well-known older gt-nonplanar proteins, e.g. methylamine dehydrogenase, some mammalian hormones such as follitropin, or most gt-nonplanar scorpion neurotoxins, were not retrieved by the procedure described, though they are in the Registry file. Such proteins are listed in the second part of Appendix II, but this list is almost certainly incomplete.

Table 4. Numbers of natural linear peptides/proteins with *c* disulfide bridges retrieved by the Registry file search (see text), and of gt-planar such proteins.

c	# proteins	gt-planar proteins	% planar
4	115	103	89.56
5	45	45	100
6	38	36	94.74
7	12	12	100
8	5	5	100
9	12	12	100
10	11	11	100
11	35	35	100
12	10	10	100
13	2	2	100
14	4	4	100
15	5	5	100
16	2	2	100
17	13	13	100
18	1	1	100
19	2	2	100
20	3	3	100
21	1	1	100
22	-	-	-
23	2	2	100

The nonplanar peptides/proteins can be classified according to their disulfide bridge patterns:

- Class 1: A large and both structurally and functionally diverse group of proteins made of about 50-65 amino acids containing 4 disulfide bridges, of both animal and plant origin, such as the well-known scorpion neurotoxins,[15,31] the sweet protein brazzein from an African fruit,[32] α-amylase inhibitors from cereal plants such as several γ-thionins, and others.[33] These all share the same nonplanar 4-bridge pattern schematically drawn in Figure 3a.

- Class 2: A small group of plant products of about 40 amino acids and 4 disulfide bridges arranged as in Figure 3b, such as pollen allergens Amb a V and Amb t V from *Ambrosia species*,[34] and PSI-1.1, a serine proteinase inhibitor from paprika.[35] Interestingly, no protein corresponding to the nonplanar patterns in Fig. 3c,d is known to us.

- Class 3: The light chain of methylamine dehydrogenase from various bacteria,[16,36] a sixfold disulfide-bridged chain of 131 amino acids with the bridges arranged as in Figure 4a. Since there is a further covalent bridge between modified tryptophanes 57 and 108 (dashed line in Figure 4a), these compounds actually contain an undecacyclic ring system (not counting

additional rings contributed by prolines).

- Class 4: The β–chains of several hormones from the vertebrate reproductive system, such as follitropin (FSH), lutropin (LH), thyrotropin (TSH), chorionic gonadotropin (CG), chains of 110 – 140 amino acids containing six disulfide bridges arranged as in Figure 4b.[37,38]

- Class 5: The plant cyclotides, a group of circular peptides of about 30 amino acids containing three disulfide bridges.[39] For a circular peptide/protein with three disulfide bridges, there are exactly one gt-nonplanar bridge pattern and exactly four planar ones. The nonplanar pattern is shown in Figure 5, top. The drawing at the left emphasizes its similarity to the linear peptides of Figure 3a, the one on the right stresses its cyclicity and serves for comparison with the planar patterns shown in Figure 5, bottom.

We did not encounter any circular peptide/protein with more than three disulfide bridges.

CONCLUSION

Based on the limited data presently available we gained the impression that Nature does not like to produce gt-nonplanar compounds. There may be a few exceptions from this rule, such as linear proteins with 4 disulfide bridges, and circular peptides with 3 disulfide bridges, in which the rate of nonplanars was found higher than expected. This however may be an artifact in our small database covering recent years only.

The synthesis of gt-nonplanar compounds seems to be inherently difficult. Synthesis of a small-size gt-nonplanar compound would require to have electron density of two covalent bonds crossing in free space (Fig. 2, **a**), or to form an extremely long and/or bent bond (**b** and **d**), or to push atoms into the interior of small rings and to manipulate them there (**c**), or to construct small cycles around atoms (**c**), no easy tasks for a chemist nor for an enzyme.

Synthesis of a macrocyclic gt-nonplanar system, on the other hand, suffers from the well-known problems of any macrocyclization. Thus, large gt-nonplanar ring systems such as Schill's **1,** Walba's Möbius strips, or Siegel's Kuratowski cyclophane form spontaneously by macrocyclization of appropriately functionalized precursors, but are always accompanied by the respective planar analogs and by products of intermolecular reactions, even under high-dilution conditions. So, small gt-nonplanar systems are hard to obtain for mostly thermodynamic reasons, large ones are hard to obtain *selectively* for reasons of kinetics. There might exist an intermediate size range more easily accessible, and some syntheses reported in the references probably profited from this fact.

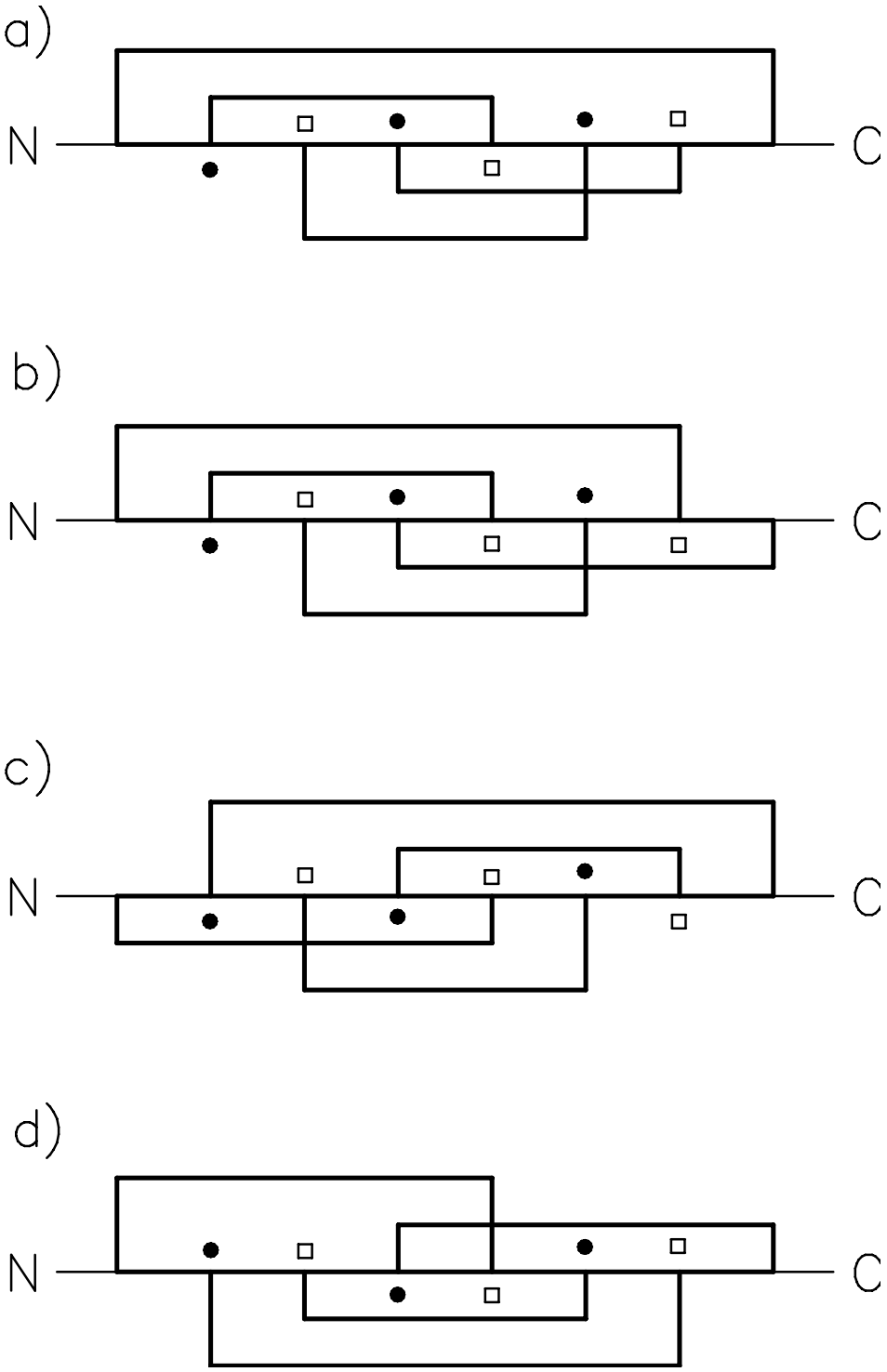

Figure 3. Disulfide-bridge patterns of gt-nonplanar proteins of 4 disulfide bridges. The polypeptide chains run from left to right, each vertex corresponds to a cysteine, lateral rectangular lines symbolize disulfide-bridges. Non-cysteine amino acids are suppressed. Note that pattern c is the reverse of pattern b, whereas patterns a and d are self-reverse (palindromic).

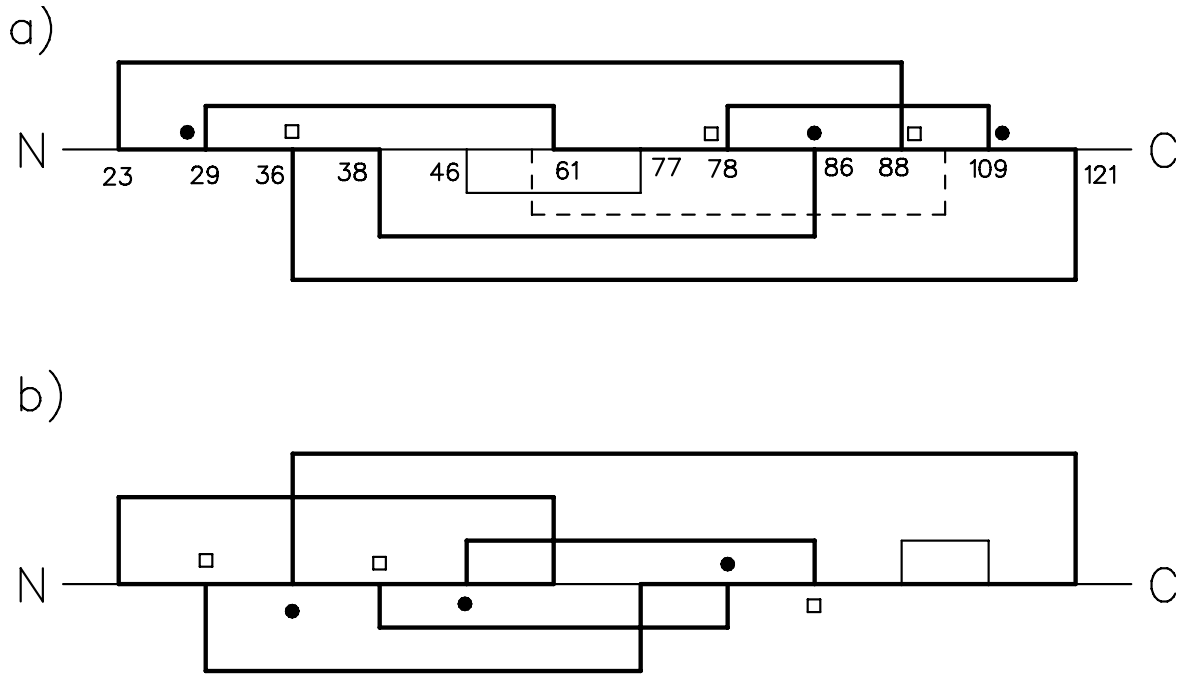

Figure 4. Disulfide-bridge patterns of known gt-nonplanar proteins of 6 disulfide bridges. The polypeptide chains run from left to right, each vertex corresponds to a cysteine, lateral rectangular lines symbolize disulfide-bridges. Non-cysteine amino acids are suppressed. For the dashed line see text.

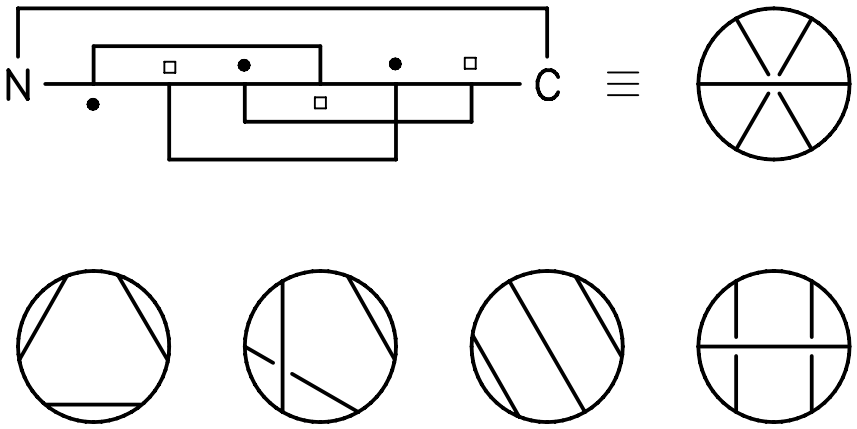

Figure 5. Disulfide-bridge patterns of circular peptides of 3 disulfide bridges. Top, the only gt-nonplanar pattern. Bottom, the four gt-planar patterns. Lines spanning across ring perimeters symbolize disulfide bridges.

gt-Nonplanar natural products, linear proteins with at least four or circular proteins with at least three disulfide bridges arranged in particular patterns, are a special case. Linear precursors may be synthesized via the normal ribosomal pathway, the macropolycyclic systems probably form on oxidation of cysteine side chains. Thus in at least one case, brazzein (54 amino acids, linear precursor synthesized on solid phase), spontaneous folding and disulfide bridge formation to the native state could be observed in detail *in vitro*.[32] In this respect gt-nonplanar brazzein is not different from similarly cysteine-rich but planar proteins.[40,41] Finally, a recent chemical synthesis of gt-nonplanar circular peptides/proteins suffered from the usual problems of macrocyclization as outlined above,[39c] while the biosynthesis of such compounds seems to be completely unknown.

Note added in Proof: We were recently informed by Dr. D. Hounshell that on testing the Beilstein compounds for gt-(non)planarity about 400 nonplanars were found. Their identities were not yet disclosed to us.

ACKNOWLEDGMENT

We thank Professor F. J. Brandenburg and Dr. M. Raitner (University of Passau) for providing the planarity test, Professor A. Lawson (Beilstein Institute, Frankfurt) for contacting MDL, and Dr. D. Hounshell and his staff (MDL, San Leandro, CA) for running the planarity test on the Beilstein compounds.

REFERENCES AND NOTES

(1) (a) Harary, F. *Graph Theory*, Addison-Wesley, Reading, Massachusetts 1969. (b) Sometimes the existence of gt-nonplanar compounds is even neglected at all, see Dury, L.; Latour, T.; Leherte, L.; Barberis, F.; Vercauteren, D. P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1437-1445. (c) It is not known whether or not gt-nonplanarity of a compound has any chemical significance. At least up to now no physical or chemical property could be demonstrated to be common to gt-nonplanar compounds and to distinguish them from the majority of (gt-planar) compounds.

(2) Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. *Fresenius J. Anal. Chem*. **2001**, *369*, 709-714.

(3) If in a complex (e.g. of the ferrocene or cryptand type) all interactions between the central metal atom and the ligand atoms are represented by lines, a multitude of gt-

nonplanar compounds is easily obtained. For a discussion of this problem see Mislow, K. *Croat. Chem. Acta* **1996**, *69*, 485-511.

(4) Simmons, H. E.; Maggio, J. E. *Tetrahedron Lett*. **1981**, *22*, 287-290. Benner, S. A.; Maggio, J. E.; Simmons, H. E. *J. Am. Chem. Soc.* **1981**, *103*, 1581-1582.

(5) Paquette, L. A., Vazeux, M. *Tetrahedron Lett*. **1981**, *22*, 291-294. Paquette, L. A.; Williams, R. V.; Vazeux, M.; Browne, A. R. *J. Org. Chem*. **1984**, *49*, 2194-2197.

(6) (a) Schill, G.; Rißler, K.; Fritz, H.; Vetter, W. *Angew. Chem. Int. Ed. Engl*. **1981**, *20*, 187. Rißler, K.; Schill, G.; Fritz, H.; Vetter, W. *Chem. Ber.* **1986**, *119*, 1374-1399. (b) Even earlier in the same research group a probably gt-nonplanar macropolycycle was obtained, however, the structure was never completely established: Schill, G.; Doerjer, G.; Logemann, E.; Fritz, H. *Chem. Ber.* **1979**, *112*, 3603-3615.

(7) Walba, D. M.; Richards, R. M.; Haltiwanger, R. C. *J. Am. Chem. Soc.* **1982**, *104*, 3219-3221. Walba, D. M. *Tetrahedron* **1985**, *41*, 3161-3212. Walba, D. M.; Armstrong, J. D.; Perry, A. E.; Richards, R. M.; Homan, T. C.; Haltiwanger, R. C. *Tetrahedron* **1986**, *42*, 1883-1894. Walba, D. M.; Homan, T. C.; Richards, R. M.; Haltiwanger, R. C. *New J. Chem*. **1993**, *17*, 661-681.

(8) Kuck, D.; Schuster, A. *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 1192. Eckrich, R.; Kuck, D. *Synlett* **1993**, 344-347. Kuck, D.; Schuster, A.; Paisdor, B.; Gestmann, D. *JCS Perkin 1* **1995**, 721-732. Kuck, D. *Synlett* **1996**, 949-965. Kuck, D. *Liebigs Ann./Recueil* **1997**, 1043-1057. Kuck, D.; Krause, R. A.; Gestmann, D.; Posteher, F.; Schuster, A. *Tetrahedron* **1998**, *54*, 5247-5258. Kuck, D. *Top. Curr. Chem*. **1998**, *196*, 167-220.

(9) Chen, C.-T.; Gantzel, P.; Siegel, J. S.; Baldridge, K. K.; English, R. B.; Ho, D. M. *Angew. Chem. Int. Ed. Engl*. **1995**, *34*, 2870.

(10) Otsubo, T.; Ogura, F.; Misumi, S. *Tetrahedron Lett*. **1983**, *24*, 4851-4854. Otsubo, T.; Aso, Y.; Ogura, F.; Misumi, S.; Kawamoto, A.; Tanaka, J. *Bull. Chem. Soc. Jpn*. **1989**, *62*, 164-170.

(11) Pinkos, R.; Melder, J.-P.; Fritz, H.; Prinzbach, H. *Angew. Chem. Int. Ed. Engl.* **1989**, *28*, 310. Pinkos, R.; Melder, J.-P.; Prinzbach, H. *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 92. Melder, J.-P.; Pinkos, R.; Fritz, H.; Wörth, J.; Prinzbach, H. *J. Am. Chem. Soc*. **1992**, *114*, 10213-10231. Pinkos, R.; Melder, J.-P.; Weber, K.; Hunkler, D.; Prinzbach, H. *J. Am. Chem. Soc*. **1993**, *115*, 7173-7191. Pinkos, R.; Weiler, A.; Voss, T.; Weber, K.; Wahl, F.; Melder, J.-P.; Fritz, H.; Hunkler, D.; Prinzbach, H. *Liebigs Ann./Recueil* **1997**, 2069-2088.

(12) Lee, C.-H.; Garcia, B.; Bruice, T. C. *J. Am. Chem. Soc.* **1990**, *112*, 6434-6435. Garcia, B.; Lee, C.-H.; Blasko, A.; Bruice, T. C. *J. Am. Chem. Soc*. **1991**, *113*, 8118-8126. Blasko, A.; Garcia, B.; Bruice, T. C. *J. Org. Chem*. **1993**, *58*, 5738-5747.

(13) Moriarty, R. M.; Rao, M. S. C.; Tuladhar, S. M.; D'Silva, C.; Williams, G.; Gilardi, R. *J. Am. Chem. Soc*. **1993**, *115*, 1194-1196. Moriarty, R. M.; Rao, M. S. C.; Tuladhar, S. M.; Gilardi, R.; D'Silva, C.; Williams, G. *J. Phys. Org. Chem*. **1994**, *7*, 303-308.

(14) Suffczynski, M. *Mol. Phys. Rep*. **1999**, *23*, 67-76.

(15) Crippen in 1974 pointed out that gt-nonplanar disulfide-bridged proteins might exist, but noticed that no such compound was known (ref. 15a). That same year two research groups published one such protein each, without recognizing its gt-nonplanarity (ref. 15b,c): (a) Crippen, G. M. *J. Theor. Biol.* **1974**, *45*, 327-338. (b) Babin, D. R.; Watt, D. D.; Goos, S. M.; Mlejnek, R. V. *Arch. Biochem. Biophys*. **1974**, *164*, 694-706. (c) Kopeyan, C.; Martinez, G.; Lissitzky, S.; Miranda, F.; Rochat, H. *Eur. J. Biochem*. **1974**, *47*, 483-489. (d) Almassy, R. J.; Fontecilla-Camps, J. C.; Suddath, F. L.; Bugg, C. E. *J. Mol. Biol*. **1983**, *170*, 497-527. (e) Fontecilla-Camps, J. C.; Habersetzer-Rochat, C.; Rochat, H. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 7443-7447.

(16) Mao, B. *J. Am. Chem. Soc.* **1989**, *111*, 6132-6136. Liang, C.; Mislow K. *J. Am. Chem. Soc*. **1994**, *116*, 3588-3592. Liang, C.; Mislow, K. *J. Am. Chem. Soc.* **1995**, *117*, 4201-4213.

(17) C. Kuratowski *Fund. Math*. **1930**, *15*, 271-283.

(18) A homeomorphic graph is a graph having a sequence of vertices (edges) where the parent graph has an edge, with the parent itself also being included. Sometimes, a homeomorphic graph is called a subdivision of its parent.

(19) Rücker, G.; Rücker, C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457-1462.

(20) Booth, K. S.; Lueker, G. S. *J. Comput. System Sci.* **1976**, *13*, 335-379.

(21) Forstner, M.; Pick, A.; Raitner, M. Graph Template Library GTL, see http://www.infosun.fmi.uni-passau.de/GTL/ .

(22) A cubic graph is a graph all vertices of which have degree 3, a quartic graph has all vertices of degree 4.

(23) Meringer, M. *J. Graph Theory* **1999**, *30*, 137-146.

(24) (a) http://www.mathe2.uni-bayreuth.de/markus/reggraphs.html. (b) Sloane's Online Encyclopedia of Integer Sequences, see http://www.research.att.com/%7Enjas/sequences.

(25) Kerber, A.; Laue, R.; Grüner, T.; Meringer, M. *MATCH – Commun. Math. Comput.*

Chem. **1998**, *37*, 205-208.

(26) All graphs of $c \leq 3$ are necessarily planar, while those of $m \geq 3n\text{-}5$ are necessarily nonplanar.[1a,17] The planarity test has no built-in knowledge of these facts but nevertheless found the correct result in all such cases.

(27) Klapper, M. H.; Klapper, I. Z. *Biochim. Biophys. Acta* **1980**, *626*, 97-105.

(28) In February 2002 the Registry file contained 75118 at least tetracyclic ring parents (ring systems, different heteroatom and multiple bond patterns being counted individually), whereas the total number of entries was roughly 35 million at that time.

(29) The search as described retrieved a number of circular peptides containing four or six cysteines. One may wonder how this can happen. The simple explanation is as follows. There is no search field for the occurrence number of a specific amino acid, rather the cysteines were searched as a sequence of eight cysteines with unspecified numbers of unspecified amino acids in between. In a circular protein the search mechanism obviously does not stop after one turn.

(30) Janssen et al. reported on a variant of phospholipase A_2 called group X PLA2 that may contain eight disulfide bridges in a nonplanar pattern: Janssen, M. J. W.; Verheij, H. M.; Slotboom, A. J.; Egmond, M. R. *Eur. J. Biochem*. **1999**, *261*, 197-207.

(31) (a) Selisko, B.; Garcia, C.; Becerril, B.; Delepierre, M.; Possani, D. *Eur. J. Biochem*. **1996**, *242*, 235-242. (b) Landon, C.; Sodano, P.; Cornet, B.; Bonmatin, J.-M.; Kopeyan, C.; Rochat, H.; Vovelle, F.; Ptak, M. *Proteins: Struct. Funct. Genet.* **1997**, *28*, 360-374. (c) Jablonsky, M. J.; Jackson, P.; Krishna, N. R. *Biochemistry* **2001**, *40*, 8273-8282.

(32) Izawa, H.; Ota, M.; Kohmura, M.; Ariyoshi, Y. *Biopolymers* **1996**, *39*, 95-101.

(33) Caldwell, J. E.; Abildgaard, F.; Dzakula, Z.; Ming, D.; Hellekant, G.; Markley, J. L. *Nature Struct. Biol*. **1998**, *5*, 427-431.

(34) Metzler, W. J.; Valentine, K.; Roebber, M.; Marsh, D. G.; Mueller, L. *Biochemistry* **1992**, *31*, 8697-8705.

(35) Antcheva, N.; Patty, A.; Athanasiadis, A.; Tchorbanov, B.; Zakhariev, S.; Pongor, S. *Proc. Eur. Pept. Symp.,* (24th), **1998** (meeting date 1996), 209-210.

(36) Chen, L.; Doi, M.; Durley, R. C. E.; Chistoserdov, A. Y.; Lidstrom, M. E.; Davidson, V. L.; Mathews, F. S. *J. Mol. Biol.* **1998**, *276*, 131-149. Sun, D.; Jones, L. H. ; Mathews, F. S. ; Davidson, V. L. *Prot. Engineering* **2001**, *14*, 675-681.

(37) Amoresano, A.; Orru, S.; Siciliano, R. A.; De Luca, E.; Napoleoni, R.; Sirna, A.; Pucci, P. *Biol. Chem*. **2001**, *382*, 961-968.

(38) Lapthorn, A. J.; Harris, D. C.; Littlejohn, A.; Lustbader, J. W.; Canfield, R. E.; Machin, K. J.; Morgan, F. J.; Isaacs, N. W. *Nature*, **1994**, *369*, 455-461.

(39) (a) Craik, D. J.; Daly, N. L.; Bond, T.; Waine, C. *J. Mol. Biol.* **1999**, *294*, 1327-1336. (b) Tam, J. P.; Lu, Y.-A.; Yang, J.-L.; Chiu, K.-W. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 8913-8918. (c) Jennings, C.; West, J.; Waine, C.; Craik, D.; Anderson, M. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10614-10619. (d) Heitz, A.; Hernandez, J.-F.; Gagnon, J.; Hong, T. T.; Pham, T. C.; Nguyen, T. M.; Le-Nguyen, D.; Chiche, L. *Biochemistry* **2001**, *40*, 7973-7983. (e) Felizmenio-Quimio, M. E.; Daly, N. L.; Craik, D. J. *J. Biol. Chem*. **2001**, *276*, 22875-22882. (f) Craik, D. J. *Toxicon* **2001**, *39*, 1809-1813.

(40) (a) Fajloun, Z.; Carlier, E.; Lecomte, C.; Geib, S.; di Luccio, E.; Bichet, D.; Mabrouk, K.; Rochat, H.; De Waard, M.; Sabatier, J.-M. *Eur. J. Biochem*. **2000**, *267*, 5149-5155. (b) di Luccio, E.; Azulay, D.-O.; Regaya, I.; Fajloun, Z.; Sandoz, G.; Mansuelle, P.; Kharrat, R.; Fathallah, M.; Carrega, L.; Estève, E.; Rochat, H.; De Waard, M.; Sabatier, J.-M. *Biochem. J.* **2001**, *358*, 681-692.

(41) Chang, J.-Y. *J. Biol. Chem*. **2002**, *277*, 120-126.

APPENDIX I. gt-Nonplanar organic compounds.

BRN	CAS-RN	Ref.	Structure
1069562	72619-62-6[a]	6b	
3572815	120497-67-8	11	**5**
3574917	120523-83-3	11	**6**
3575191	124316-48-9	11	
3575192	124316-63-8	11	
4173759	129031-63-6	12	**3**
4415047	77973-29-6	4,5	
4574897	88903-18-8	10	
4610048	345242-79-7[b]	10	
4771833	135005-80-0	8	
4771834	135005-81-1	8	
5218180	81617-30-3[c]	7	
5228553	108801-75-8	7	
5262360	see 4415047		
5628262	-	11	
5646070	120497-73-6	11	
5647263	120497-66-7	11	
5660006	116303-24-3	8	
5889340	146698-08-0	13	**2** *k=5*
5891882	146698-09-1	13	**2** *k=6*
5893191	145292-80-4	11	**4**
5899223	146698-06-8	13	
6364798	151419-97-5[d]	11	
6471846	100084-23-9	6a	**1**
6830315	see 5660006		
6831464	174782-93-5	8	
6937420	157766-33-1	8	
6940370	146698-07-9	13	**2** *k=4*
6942468	see 5889340		
6942687	157766-31-9	8	
6942689	157766-32-0	8	
6944439	see 5891882		
6949145	157766-35-3	8	
6949169	157766-34-2	8	
6950636	see 5899223		
7102893	see 4574897		
7233909	see 5660006		
7450090	see 6831464		
7450091	174782-95-7	8	
7534708	182071-99-4	8	
7541512	182071-98-3	8	
7543620	-	8	
7547652	-	8	
7552840	182071-95-0	8	
7553617	182071-97-2	8	
7553639	182071-96-1	8	
7555036	182071-94-9	8	
7566443	174619-33-1	9	
7773170	197573-36-7	11	
7779454	197573-37-8	11	
7780356	see 3572815		
7781205	197573-34-5	11	
7782177	see 3574917		
7782666	197573-28-7	11	
8173869	208470-79-5	8	
8175888	208470-83-1	8	
8176598	208470-82-0	8	
8177211	208470-81-9	8	

[a] also 343827-97-4. [b] also 88931-19-5, 88903-16-6, 88903-15-5.
[c] also 141272-33-2, 141271-34-3, 141271-35-4. [d] also 151419-96-4.

APPENDIX II. Part 1: Recent gt-nonplanar peptides/proteins obtained by searching the Registry file (see text).

CAS-RN	name	CAS-RN	name
363609-48-7	Bs-dprlT1	256447-30-0	Cycloviolin C
278596-43-3	MCoTI-II	256447-29-7	Cycloviolin B
262854-46-6	BmK M2	256447-28-6	Cycloviolin A
261714-39-0	Kalata S	245764-58-3	
261714-38-9	Kalata B4	220326-07-8	BmK AS-1
261714-37-8	Kalata B3	216101-47-2	PSI-1.1
261714-36-7	Kalata B2	198085-36-8	Drosomycin
261714-35-6	Kalata B5	187887-84-9	Cn10, ref 31a
261714-34-5	Cycloviolacin O12	186101-24-6	mutated CG-β
261714-33-4	Cycloviolacin H1	186101-23-5	mutated CG-β
261714-32-3	Cycloviolacin O11	186101-22-4	mutated CG-β
261714-31-2	Cycloviolacin O10	186101-21-3	mutated CG-β
261714-30-1	Cycloviolacin O9	186101-20-2	mutated CG-β
261714-29-8	Cycloviolacin O8	182578-08-1	ostrich LSH-β
261714-28-7	Cycloviolacin O7	182578-07-0	ostrich FSH-β
261714-27-6	Cycloviolacin O6	178359-31-4	enantio-Brazzein
261714-26-5	Cycloviolacin O5	175069-47-3	Brazzein
261714-25-4	Cycloviolacin O4	172308-69-9	ω-Hordothionin
261714-24-3	Cycloviolacin O3	165587-35-9	γ-Thionin
261714-23-2	Cycloviolacin O2	165587-32-6	γ-Thionin
261714-22-1	Cycloviolacin O1	165587-29-1	γ-Thionin
259853-19-5	Circulin F	161471-68-7	Cyclopsychotride A
259853-18-4	Circulin E	153421-82-0	CsE 1
259853-17-3	Circulin D	143063-12-1	Allergen Amb a V
259853-16-2	Circulin C		Allergen Amb t V
256447-31-1	Cycloviolin D	59473-22-2	Kalata B1

Part 2: gt-Nonplanar peptides/proteins published earlier.

CAS-RN	name
76415-77-5	CsE v3, ref 15
	AaH toxin II, ref 15
	Lqq toxin III, ref 31
60496-14-2	MADH (light chain), ref 36
9002-68-0	Follitropin, FSH (β-chain), ref 37
9002-61-3	Chorionic Gonadotropin, CG (β-chain), ref 38